Forecast Evaluation

Mingmian Cheng, Norman R. Swanson and Chun Yao

Abstract The development of new tests and methods used in the evaluation of time series forecasts and forecasting models remains as important today as it has for the last 50 years. Paraphrasing what Sir Clive W.J. Granger (arguably the father of modern day time series forecasting) said in the 1990s at a conference in Svinkloev, Denmark, 'OK, the model looks like an interesting extension, but can it forecast better than existing models.' Indeed, the forecast evaluation literature continues to expand, with interesting new tests and methods being developed at a rapid pace. In this chapter, we discuss a select a group of predictive accuracy tests and model selection methods that have been developed in recent years, and that are now widely used in the forecasting literature. We begin by reviewing several tests for comparing the relative forecast accuracy of different models, in the case of point forecasts. We then broaden the scope of our discussion by introducing density-based predictive accuracy tests. We conclude by noting that predictive accuracy is typically assessed in terms of a given loss function, such as mean squared forecast error or mean absolute forecast error. Most tests, including those discussed here, are consequently loss function dependent, and the relative forecast superiority of predictive models is therefore also dependent on specification of a loss function. In light of this fact, we conclude this chapter by discussing loss function robust predictive density accuracy tests that have recently been developed using principles of stochastic dominance.

Mingmian Cheng
Department of Finance, Lingnan (University) College, Sun Yat-sen University, 135 Xingang West Road, Guangzhou, China 510275, e-mail: chengmm3@mail.sysu.edu.cn

Norman R. Swanson
Department of Economics, School of Arts and Sciences, Rutgers University, 75 Hamilton Street, New Brunswick, NJ, USA 08901, e-mail: nswanson@economics.rutgers.edu

Chun Yao
Department of Economics, School of Arts and Sciences, Rutgers University, 75 Hamilton Street, New Brunswick, NJ, USA 08901, e-mail: cyao@economics.rutgers.edu

Part I: Forecast Evaluation Using Point Predictive Accuracy Tests

In this section, our objective is to review various commonly used statistical tests for comparing the relative accuracy of point predictions from different econometric models. Four main groups of tests are outlined: (i) tests for comparing two non-nested models, (ii) tests for comparing two nested models, (iii) tests for comparing multiple models, where at least one model is non-nested, and (iv) tests that are consistent against generic alternative models. The papers cited in this section (and in subsequent sections) contain references to a large number of papers that develop alternative related tests.

Of note is that the tests that we discuss in the sequel assume that all competing models are approximations to some unknown underlying data generating process, and are thus potentially misspecified. The objective, this is to select the "best" model from amongst multiple alternatives, where "best" refers to a given loss function, say.

1 Comparison of two non-nested models

The starting point of our discussion is the Diebold-Mariano (DM: [1]) test for the null hypothesis of equal predictive accuracy between two competing models, given a pre-specified loss function. This test sets the groundwork for many subsequent predictive accuracy tests. The DM test assumes that parameter estimation error is asymptotically negligible by positing that the number of observations used for in-sample model estimation grows faster than the number of observations used in out-of-sample forecast evaluation. Parameter estimation error in DM tests, which are often also called DM-West tests, is explicitly taken into account of in [2], although at the cost of requiring that the loss function is differentiable.

To fix ideas and notation, let $u_{i,t+h} = y_{t+h} - f_i(Z_i^t, \theta_i^\dagger)$ be the h-step ahead forecast error associated with the i-th model, $f_i(\cdot, \theta_i^\dagger)$, where the benchmark model is always denoted as "model 0", i.e. $f_0(\cdot, \theta_0^\dagger)$. As $\theta_i^\dagger$ and thus $u_{i,t+h}$ are unknown, we construct test statistics using $\widehat{\theta}_{i,t}$ and $\widehat{u}_{i,t+h} = y_{t+h} - f_i(Z_i^t, \widehat{\theta}_{i,t})$, where $\widehat{\theta}_{i,t}$ is an estimator of $\theta_i^\dagger$ constructed using information in Z_i^t from time periods 1 to t, under a recursive estimation scheme, or from $t-R+1$ to t, under a rolling-window estimation scheme. Hereafter, for notional simplicity, we only consider the recursive estimation scheme, and the rolling-window estimation scheme can be treated in an analogous manner. To do this, split the total sample of T observations into two sub-samples of length R and n, i.e. $T = R + n$, where only the last n observations are used for forecast evaluation. At each step, we first estimate the model parameters as follows,

$$\widehat{\theta}_{i,t} = \arg\min_{\theta_i} \frac{1}{t} \sum_{j=1}^{t} q(y_j - f_i(Z_i^{j-1}, \theta_i)), \quad t \geq R \tag{1}$$

These parameters are used to parameterize the prediction model, and an h-step-ahead prediction (and prediction error) is constructed. This procedure is repeated by adding one new observation to the original sample, yielding a new h-step-ahead prediction (and prediction error). In such a manner, we can construct a sequence of $(n-h+1)$ h-step ahead prediction errors. For a given loss function, $g(\cdot)$, the null hypothesis of DM test is specified as,

$$H_0 : E(g(u_{0,t+h}) - g(u_{1,t+h})) = 0$$

against

$$H_A : E(g(u_{0,t+h}) - g(u_{1,t+h})) \neq 0$$

Of particular note here is that the loss function $g(\cdot)$ used for forecast evaluation may not be the same as the loss function $q(\cdot)$ used for model estimation in Equation (1). However, if they are the same (e.g. models are estimated by ordinary least square (*OLS*) and forecasts are evaluated by a quadratic loss function, say), parameter estimation error is asymptotically negligible, regardless of the limiting ratio of n/R, as $T \to \infty$.

Define the following statistic,

$$\widehat{S}_n(0,1) = \frac{1}{\sqrt{n}} \sum_{t=R-h+1}^{T-h} (g(\widehat{u}_{0,t+h}) - g(\widehat{u}_{1,t+h}))$$

then,

$$\begin{aligned} \widehat{S}_n(0,1) - S_n(0,1) = {} & E(\nabla_{\theta_0} g(u_{0,t+h})) \frac{1}{\sqrt{n}} \sum_{t=R-h+1}^{T-h} (\widehat{\theta}_{0,t+h} - \theta_0^{\dagger}) \\ & - E(\nabla_{\theta_1} g(u_{1,t+h})) \frac{1}{\sqrt{n}} \sum_{t=R-h+1}^{T-h} (\widehat{\theta}_{1,t+h} - \theta_1^{\dagger}) + o_p(1) \end{aligned} \tag{2}$$

The limiting distribution of the right-hand side of Equation (2) is given by Lemma 4.1 and Theorem 4.1 in [2]. From Equation (2), we can immediately see that if $g(\cdot) = q(\cdot)$, then $E(\nabla_{\theta_i} g(u_{i,t+h})) = 0$ by the first order conditions, and parameter estimation error is asymptotically negligible. Another situation in which parameter estimation error vanishes asymptotically is when $n/R \to 0$, as $T \to \infty$.

Without loss of generality, consider the case of $h = 1$. All results carry over to the case when $h > 1$. The DM test statistic is given by,

$$\widehat{DM}_n = \frac{1}{\sqrt{n}} \frac{1}{\widehat{\sigma}_n} \sum_{t=R}^{T-1} (g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{1,t+1}))$$

with

$$\begin{aligned}\widehat{\sigma}_n =& \widehat{S}_{gg} + 2\Pi\widehat{F}_0'\widehat{A}_0\widehat{S}_{h_0h_0} + 2\Pi\widehat{F}_1'\widehat{A}_1\widehat{S}_{h_1h_1}\widehat{A}_1\widehat{F}_1 \\ & - 2\Pi(\widehat{F}_1'\widehat{A}_1\widehat{S}_{h_1h_0}\widehat{A}_0\widehat{F}_0 + \widehat{F}_0'\widehat{A}_0\widehat{S}_{h_0h_1}\widehat{A}_1\widehat{F}_1) \\ & + \Pi(\widehat{S}_{gh_1}'\widehat{A}_1\widehat{F}_1 + \widehat{F}_1'\widehat{A}_1\widehat{S}_{gh_1})\end{aligned}$$

where for $i, j = 0, 1$, $\Pi = 1 - \frac{\ln(1+\pi)}{\pi}$, and $q_t(\widehat{\theta}_{i,t}) = q(y_t - f_i(Z_i^{t-1}, \widehat{\theta}_{i,t}))$,

$$\widehat{S}_{h_ih_j} = \frac{1}{n}\sum_{\tau=-l_n}^{l_n} w_\tau \sum_{t=R+l_n}^{T-l_n} \nabla_\theta q_t(\widehat{\theta}_{i,t})\nabla_\theta q_{t+\tau}(\widehat{\theta}_{j,t})'$$

$$\begin{aligned}\widehat{S}_{gh_i} =& \frac{1}{n}\sum_{\tau=-l_n}^{l_n} w_\tau \sum_{t=R+l_n}^{T-l_n} \left(g(\widehat{u}_{0,t}) - g(\widehat{u}_{1,t}) - \frac{1}{n}\sum_{t=R}^{T-1}(g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{1,t+1}))\right) \\ & \times \nabla_\theta q_{t+\tau}(\widehat{\theta}_{i,t})' \\ \widehat{S}_{gg} =& \frac{1}{n}\sum_{\tau=-l_n}^{l_n} w_\tau \sum_{t=R+l_n}^{T-l_n} \left(g(\widehat{u}_{0,t}) - g(\widehat{u}_{1,t}) - \frac{1}{n}\sum_{t=R}^{T-1}(g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{1,t+1}))\right) \\ & \times \left(g(\widehat{u}_{0,t+\tau}) - g(\widehat{u}_{1,t+\tau}) - \frac{1}{n}\sum_{t=R}^{T-1}(g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{1,t+1}))\right)\end{aligned}$$

with $w_\tau = 1 - \frac{\tau}{l_n - 1}$, and

$$\widehat{F}_i = \frac{1}{n}\sum_{t=R}^{T-1} \nabla_{\theta_i} g(\widehat{u}_{i,t+1}), \qquad \widehat{A}_i = \left(-\frac{1}{n}\sum_{t=R}^{T-1} \nabla_{\theta_i}^2 q(\widehat{\theta}_{i,t})\right)^{-1}$$

Assumption 1.1: (y_t, Z^{t-1}), with y_t scalar and Z^{t-1} an $\Re^\zeta$-valued $(0 < \zeta < \infty)$ vector, is a strictly stationary and absolutely regular β-mixing process with size $-4(4+\psi)/\psi$, $\psi > 0$.

Assumption 1.2: (i) $\theta^\dagger$ is uniquely identified (i.e. $E(q(y_t, Z^{t-1}, \theta_i))) > E(q(y_t, Z^{t-1}, \theta_i^\dagger)))$ for any $\theta_i \neq \theta_i^\dagger$); (ii) $q(\cdot)$ is twice continuously differentiable on the interior of Θ, and for Θ a compact subset of $\Re^\rho$; (iii) the elements of $\nabla_\theta q$ and $\nabla_\theta^2 q$ are p-dominated on Θ, with $p > 2(2+\psi)$, where ψ is the same positive constant as defined in Assumption 1.1; and (iv) $E(-\nabla_\theta^2 q)$ is negatively definite uniformly on Θ.

PROPOSITION 1.1 (From Theorem 4.1 in [2]): With Assumptions 1.1 and 1.2, also, assume that $g(\cdot)$ is continuously differentiable, then, if as $n \to \infty$, $l_n \to \infty$ and $l_n/n^{1/4} \to 0$, then as $n,\ R \to \infty$, under H_0,

$$\widehat{DM}_n \xrightarrow{d} N(0,1)$$

Under H_A,

$$\Pr(n^{-1/2}|\widehat{DM}_n| > \varepsilon) \to 1, \quad \forall \varepsilon > 0$$

It is immediate to see that if either $g(\cdot) = q(\cdot)$ or $n/R \to 0$, as $T \to \infty$, the estimator $\widehat{\sigma}_n$ collapses to $\widehat{S}_{gg}$. Note that the limiting distribution of DM test obtains only for the case of short-memory series. [3] extends the DM test to the case of cointegrated variables and [4] to the case of series with high persistence. Finally, note that the two competing models are assumed to be non-nested. If they are nested, then $u_{0,t+h} = u_{1,t+h}$ under the null, and both $\sum_{t=R-h+1}^{T-h}(g(\widehat{u}_{0,t+h}) - g(\widehat{u}_{1,t+h}))$ and $\widehat{\sigma}_n$ converge in probability to zero at the same rate if $n/R \to 0$. Therefore the DM test statistic does not converge in distribution to a standard normal variable under the null. Comparison of nested models is introduced in the next section.

2 Comparison of two nested models

There are situations in which we may be interested in comparing forecasts from nested models. For instance, one of the driving forces behind the literature on out-of-sample comparison of nested models is the seminal paper by [5], who find that no models driven by economic fundamentals can beat a simple random walk model, in terms of out-of-sample predictive accuracy, when forecasting exchange rates. The models studied in this paper are nested, in the sense that parameter restrictions can be placed on the more general models that reduce these models to the random walk benchmark studied by these authors. When testing out-of-sample Granger causality, alternative models are also nested. Since the DM test discussed above is valid only when the competing models are non-nested, we introduce alternative tests that address testing among nested models.

2.1 Clark and McCracken tests for nested models

[6] (CMa) and [7] (CMb) propose several tests for nested linear models, under the assumption that prediction errors follow martingale difference sequences (this rules out the possibility of dynamic misspecification under the null for these particular tests), where CMa tests are tailored for the case of one-step-ahead forecasts, and CMb tests for the case of multi-step-ahead forecasts.

Consider the following two nested models. The restricted model is,

$$y_t = \sum_{j=1}^{q} \beta_j y_{t-j} + \varepsilon_t$$

and the unrestricted model is,

$$y_t = \sum_{j=1}^{q} \beta_j y_{t-j} + \sum_{j=1}^{k} \alpha_j x_{t-j} + u_t \tag{3}$$

The null hypothesis of CMa tests is formulated as,

$$H_0 : E(\varepsilon_t^2) - E(u_t^2) = 0$$

against

$$H_A : E(\varepsilon_t^2) - E(u_t^2) > 0$$

We can immediately see from the null and the alternative hypotheses that CMa tests implicitly assume that the restricted model cannot beat the unrestricted model. This is the case when the models are estimated by *OLS* and the quadratic loss function is employed for evaluation.

CMa propose the following three different test statistics,

$$ENC-T = (n-1)^{1/2} \frac{\overline{c}}{(n^{-1}\sum_{t=R}^{T-1}(c_{t+1}-\overline{c}))^{1/2}}$$

$$ENC-REG = (n-1)^{1/2} \frac{n^{-1}\sum_{t=R}^{T-1}(\widehat{\varepsilon}_{t+1}(\widehat{\varepsilon}_{t+1}-\widehat{u}_{t+1}))}{(n^{-1}\sum_{t=R}^{T-1}(\widehat{\varepsilon}_{t+1}-\widehat{u}_{t+1})^2 n^{-1}\sum_{t=R}^{T-1}\widehat{\varepsilon}_{t+1}^2 - \overline{c})^{1/2}}$$

$$ENC-NEW = n\frac{\overline{c}}{n^{-1}\sum_{t=1}\widehat{u}_{t+1}^2}$$

where $c_{t+1} = \widehat{\varepsilon}_{t+1}(\widehat{\varepsilon}_{t+1} - \widehat{u}_{t+1})$, $\overline{c} = n^{-1}\sum_{t=R}^{T-1} c_{t+1}$, and $\widehat{\varepsilon}_{t+1}$ and $\widehat{u}_{t+1}$ are *OLS* residuals.

Assumption 2.1: (y_t, x_t) are strictly stationary and strong mixing processes, with size $\frac{-4(4+\delta)}{\delta}$, for some $\delta > 0$, and $E(y_t^8)$ and $E(x_t^8)$ are both finite.

Assumption 2.2: Let $z_t = (y_{t-1}, ..., y_{t-q}, x_{t-1}, ..., x_{t-q})$ and $E(z_t u_t | \mathscr{F}_{t-1}) = 0$, where $\mathscr{F}_{t-1}$ is the σ-field up to time $t-1$, generated by $(y_{t-1}, y_{t-2}, ..., x_{t-1}, x_{t-2}, ...)$. Also, $E(u_t^2 | \mathscr{F}_{t-1}) = \sigma_u^2$.

Note that Assumption 2.2 assumes that the unrestricted model is dynamically correct and that u_t is conditionally homoskedastic.

PROPOSITION 2.1 (From Theorem 3.1–3.3 in [6]): With Assumptions 2.1 and 2.2, under the null, (i) if as $T \to \infty$, $n/R \to \pi > 0$, then $ENC-T$ and $ENC-REG$ converge in distribution to Γ_1/Γ_2 where $\Gamma_1 = \int_{(1+\pi)^{-1}}^{1} s^{-1} W_s' dW_s$ and $\Gamma_2 = \int_{(1+\pi)^{-1}}^{1} W_s' W_s ds$, with W_s a k-dimensional standard Brownian motion (here k is the number of restrictions or the number of extra regressors in the unrestricted model). $ENC-NEW$ converges in distribution to Γ_1. (ii) If as $T \to \infty$, $n/R \to 0$, then $ENC-T$ and $ENC-REG$ converge in distribution to $N(0,1)$. $ENC-NEW$

converges in probability to 0.

Therefore, as $T \to \infty$ and $n/R \to \pi > 0$, all three test statistics have non-standard limiting distributions. Critical values are tabulated for different k and π in CMa. Also note that the above proposition is valid only when $h = 1$, i.e. the case of one-step ahead forecasts, since Assumption 2.2 is violated when $h > 1$. For this case, CMb propose a modified test statistic for which $MA(h-1)$ errors are allowed. Namely, they propose using the following statistic:

$$ENC-T^{'} = (n-h+1)^{1/2} \times \frac{(n-h+1)^{-1}\sum_{t=R}^{T-h}\widehat{c}_{t+h}}{((n-h+1)^{-1}\sum_{j=-\overline{j}}^{\overline{j}}\sum_{t=R+j}^{T-h}K(\frac{j}{M})(\widehat{c}_{t+h}-\overline{c})(\widehat{c}_{t+h-j}-\overline{c}))^{1/2}},$$

where $K(\cdot)$ is a kernel and $0 \leq K(\frac{j}{M}) \leq 1$, with $K(0) = 1$ and $M = o(n^{1/2})$, and $\overline{j}$ does not grow with the sample size. Therefore, the denominator of $ENC-T^{'}$ is a consistent estimator of the long-run variance when $E(c_t c_{t+|k|}) = 0$ for all $|k| > h$. Of particular note is that although $ENC-T^{'}$ allows for $MA(h-1)$ errors, dynamic misspecification under the null is still not allowed. Also note that, when $h = 1$, $ENC-T^{'}$ is equivalent to $ENC-T$.

Another test statistic suggested in CMb is a DM-type test with nonstandard critical values that are needed in order to modify the DM test in order to allow for the comparison of nested models. The test statistic is:

$$MSE-T^{'} = (n-h+1)^{1/2} \times \frac{(n-h+1)^{-1}\sum_{t=R}^{T-h}\widehat{d}_{t+h}}{((n-h+1)^{-1}\sum_{j=-\overline{j}}^{\overline{j}}\sum_{t=R+j}^{T-h}K(\frac{j}{M})(\widehat{d}_{t+h}-\overline{d})(\widehat{d}_{t+h-j}-\overline{d}))^{1/2}}$$

where $\widehat{d}_{t+h} = \widehat{u}^2_{t+h} - \widehat{\varepsilon}^2_{t+h}$ and $\overline{d} = (n-h+1)^{-1}\sum_{t=R}^{T-h}\widehat{d}_{t+h}$.

Evidently, this test is a standard DM test, although it should be stressed that the critical values used in the application of this variant of the test are different. The limiting distributions of the $ENC-T^{'}$ and $MSE-T^{'}$ are provided in CMb, and are non-standard. Moreover, for the case of $h > 1$, the limiting distributions contain nuisance parameters, so that critical values cannot be tabulated directly. Instead, CMb suggest a modified version of the bootstrap method in [8] to carry out statistical inference. For this test, the block bootstrap can also be used to carry out inference (see [11] for details.)

2.2 Out-of-sample tests for Granger causality

CMa and CMb tests do not take dynamic misspecification into account under the null. [9] (CCS) propose out-of-sample tests for Granger causality allowing for pos-

sible dynamic misspecification and conditional heteroskedascity. The idea is very simple. If the coefficients $\alpha_j, j = 1,...,k$ in Equation (3) are all zeros, then residuals ε_{t+1} are uncorrelated with lags of x. As a result, including regressors $x_{t-j}, j = 1,...,k$ does not help improve predictive accuracy, and the unrestricted model does not outperform the restricted model.

Hereafter, for notational simplicity, we only consider the case of $h = 1$. All results can be generalized to the case of $h > 1$. Formally, the test statistic is,

$$m_n = n^{-1/2} \sum_{t=R}^{T-1} \widehat{\varepsilon}_{t+1} X_t,$$

where $X_t = (x_t, x_{t-1}, ..., x_{t-k-1})'$. The null hypothesis and the alternative hypothesis are formulated as,

$$H_0 : E(\varepsilon_{t+1} x_{t-j}) = 0, \quad j = 0, 1, ..., k-1$$

$$H_A : E(\varepsilon_{t+1} x_{t-j}) \neq 0, \quad \text{for some } j$$

Assumption 2.3: (y_t, x_t) are strictly stationary and strong mixing processes, with size $\frac{-4(4+\delta)}{\delta}$, for some $\delta > 0$, and $E(y_t^8)$ and $E(x_t^8)$ are both finite. $E(\varepsilon_t y_{t-j}) = 0, j = 1, 2, ..., q$.

PROPOSITION 2.2 (From Theorem 1 in [9]): With Assumption 2.3, as $T \to \infty$, $n/R \to \pi, 0 \leq \pi < \infty$, (i) under the null, for $0 < \pi < \infty$,

$$m_n \xrightarrow{d} N(0, \Xi)$$

with

$$\begin{aligned} \Xi =& S_{11} + 2(1 - \pi^{-1} \ln(1+\pi)) F' M S_{22} M F - \\ & (1 - \pi^{-1} \ln(1+\pi))(F' M S_{12} + S_{12}' M F) \end{aligned}$$

where $F = E(Y_t X_t')$, $M = \text{plim}(\frac{1}{t} \sum_{j=q}^{t} Y_j Y_j')^{-1}$, and $Y_j = (y_{j-1}, ..., y_{j-q})'$. Furthermore,

$$S_{11} = \sum_{j=-\infty}^{\infty} E((X_t \varepsilon_{t+1} - \mu)(X_{t-j} \varepsilon_{t-j+1} - \mu)')$$

$$S_{22} = \sum_{j=-\infty}^{\infty} E((Y_{t-1} \varepsilon_t)(Y_{t-j-1} \varepsilon_{t-j})')$$

$$S_{12} = \sum_{j=-\infty}^{\infty} E((\varepsilon_{t+1} X_t - \mu)(Y_{t-j-1} \varepsilon_{t-j})')$$

where $\mu = E(X_t \varepsilon_{t+1})$. In addition, for $\pi = 0$,

$$m_n \xrightarrow{d} N(0, S_{11})$$

(ii) Under the alternative,

$$\lim_{n\to\infty} \Pr(|n^{-1/2} m_n| > 0) = 1$$

COROLLARY 2.1 (From Corollary 2 in [9]): With Assumption 2.3, as $T \to \infty$, $n/R \to \pi, 0 \leq \pi < \infty, l_T \to \infty, l_T/T^{1/4} \to 0$, (i) under the null, for $0 < \pi < \infty$,

$$m_n' \widehat{\Xi}^{-1} m_n \xrightarrow{d} \chi^2_k$$

with

$$\begin{aligned}\widehat{\Xi} = \widehat{S}_{11} &+ 2(1 - \pi^{-1}\ln(1+\pi))\widehat{F}'\widehat{M}\widehat{S}_{22}\widehat{M}\widehat{F} \\ &- (1 - \pi^{-1}\ln(1+\pi))(\widehat{F}'\widehat{M}\widehat{S}_{12} + \widehat{S}_{12}'\widehat{M}\widehat{F})\end{aligned}$$

where $\widehat{F} = n^{-1}\sum_{t=R}^{T} Y_t X_t'$, $\widehat{M} = (n^{-1}\sum_{t=R}^{T-1} Y_t Y_t')r^{-1}$, and

$$\begin{aligned}\widehat{S}_{11} =& \frac{1}{n}\sum_{t=R}^{T-1}(\widehat{\varepsilon}_{t+1}X_t - \widehat{\mu}_1)(\widehat{\varepsilon}_{t+1}X_t - \widehat{\mu}_1)' \\ &+ \frac{1}{n}\sum_{t=\tau}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1}(\widehat{\varepsilon}_{t+1}X_t - \widehat{\mu}_1)(\widehat{\varepsilon}_{t+1-\tau}X_{t-\tau} - \widehat{\mu}_1)' \\ &+ \frac{1}{n}\sum_{t=\tau}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1}(\widehat{\varepsilon}_{t+1-\tau}X_{t-\tau} - \widehat{\mu}_1)(\widehat{\varepsilon}_{t+1}X_t - \widehat{\mu}_1)'\end{aligned}$$

$$\begin{aligned}\widehat{S}_{12} =& \frac{1}{n}\sum_{\tau=0}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1}(\widehat{\varepsilon}_{t+1-\tau}X_{t-\tau} - \widehat{\mu}_1)(Y_{t-1}\widehat{\varepsilon}_t)' \\ &+ \frac{1}{n}\sum_{\tau=1}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1}(\widehat{\varepsilon}_{t+1}X_t - \widehat{\mu}_1)(Y_{t-1-\tau}\widehat{\varepsilon}_{t-\tau})'\end{aligned}$$

$$\begin{aligned}\widehat{S}_{22} =& \frac{1}{n}\sum_{t=R}^{T-1}(Y_{t-1}\widehat{\varepsilon}_t)(Y_{t-1}\widehat{\varepsilon}_t)' \\ &+ \frac{1}{n}\sum_{\tau=1}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1}(Y_{t-1}\widehat{\varepsilon}_t)(Y_{t-1-\tau}\widehat{\varepsilon}_{t-\tau})' \\ &+ \frac{1}{n}\sum_{\tau=1}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1}(Y_{t-1-\tau}\widehat{\varepsilon}_{t-\tau})(Y_{t-1}\widehat{\varepsilon}_t)'\end{aligned}$$

with $w_\tau = 1 - \frac{\tau}{l_T+1}$. In addition, for $\pi = 0$,

$$m_n^{'}\widehat{S}_{11}^{-1}m_n \xrightarrow{d} \chi_k^2$$

(ii) Under the alternative, $m_n^{'}\widehat{S}_{11}^{-1}m_n$ diverges at rate n.

Note that a "nonlinear" variant of the above CCS test has also been developed by the same authors. In this generic form of the test, one can test for nonlinear Granger causality, for example, where the alternative hypothesis is that some (unknown) function of the x_t can be added to the benchmark linear model that contains no x_t in order to improve predictive accuracy. This alternative test is thus consistent against generic nonlinear alternatives. Complete details of this test are given in the next section.

3 A predictive accuracy test that is consistent against generic alternatives

The test discussed in the previous subsection is designed to have power against a given (linear) alternative; and while it may have power against other alternatives, it is not designed to do so. Thus, it is not consistent against generic alternatives. Tests that are consistent against generic alternatives are sometimes called portmanteau tests, and it is this sort of extension of the out-of-sample Granger causality test discussed above that we now turn our attention to. Broadly speaking, the above consistency has been studied in the consistent specification testing literature (see [18], [19], [20], [21], [22] and [23]).

[24] draw on both the integrated conditional moment (ICM) testing literature of [18] and [19] and on the predictive accuracy testing literature; and propose an out-of-sample version of the ICM test that is consistent against generic nonlinear alternatives. This test is designed to examine whether there exists an unknown (possibly nonlinear) alternative model with better predictive power than the benchmark model, for a given loss function. A typical example is the case in which the benchmark model is a simple autoregressive model and we want to know whether including some unknown functions of the past information can produce more accurate forecasts. This is the case of nonlinear Granger causality testing discussed above. Needless to say, this test can be applied to many other cases. One important feature of this test is that the same loss function is used for in-sample model estimation and out-of-sample predictive evaluation (see [25] and [26]).

Consider the following benchmark model,

$$y_t = \theta_1^{\dagger} y_{t-1} + u_t,$$

where $\theta_1^{\dagger} = \arg\min_{\theta_1 \in \Theta_1} E(q(y_t - \theta_1 y_{t-1}))$. The generic alternative model is,

$$y_t = \theta_{2,1}^{\dagger}(\gamma) y_{t-1} + \theta_{2,2}^{\dagger}(\gamma)\omega(Z^{t-1}, \gamma) + \upsilon_t$$

where $\theta_2^{\dagger}(\gamma) = (\theta_{2,1}^{\dagger}(\gamma), \theta_{2,2}^{\dagger}(\gamma))' = \arg\min_{\theta_2 \in \Theta_2} E(q(y_t - \theta_{2,1}(\gamma) y_{t-1} - \theta_{2,2}(\gamma)\omega(Z^{t-1}, \gamma)))$. The alternative model is "generic" due to the term $\omega(Z^{t-1}, \gamma)$, where the function $\omega(\cdot)$ is a generically comprehensive function, as defined in [18] and [19]. The test hypotheses are:

$$H_0 : E(g(u_t) - g(\upsilon_t)) = 0$$

$$H_A : E(g(u_t) - g(\upsilon_t)) > 0$$

By definition, it is clear that the benchmark model is nested within the alternative model. Thus the former model can never outperform the latter. Equivalently, the hypotheses can be restated as,

$$H_0 : \theta_{2,2}^{\dagger}(\gamma) = 0$$

$$H_A : \theta_{2,2}^{\dagger}(\gamma) \neq 0$$

Note that, given the definition of $\theta_2^{\dagger}(\gamma)$, we have that

$$E\left(g'(\upsilon_t) \times \left(-y_t, -\omega(Z^{t-1}, \gamma)\right)'\right) = 0$$

Hence, under the null, we have that $\theta_{2,2}^{\dagger}(\gamma) = 0$, $\theta_{2,1}^{\dagger}(\gamma) = \theta_1^{\dagger}$ and $E(g'(u_t)\omega(Z^{t-1}, \gamma)) = 0$. As a result, the hypotheses can be once again be restated as,

$$H_0 : E(g'(u_t)\omega(Z^{t-1}, \gamma)) = 0$$

$$H_A : E(g'(u_t)\omega(Z^{t-1}, \gamma)) \neq 0$$

The test statistic is given by

$$M_n = \int m_n(\gamma)^2 \phi(\gamma) d\gamma$$

with

$$m_n(\gamma) = n^{-1/2} \sum_{t=R}^{T-1} g'(\widehat{u}_t + 1)\omega(Z^t, \gamma)$$

where $\int \phi(\gamma) d\gamma = 1$, $\phi(\gamma) \geq 0$, and $\phi(\gamma)$ is absolutely continuous with respect to Lebesgue measure.

Assumption 4.1: (i) (y_t, Z^t) is a strictly stationary and absolutely regular strong mixing sequence with size $-4(4+\psi)/\psi, \psi > 0$; (ii) $g(\cdot)$ is three times continuously differentiable in θ, over the interior of Θ, and $\nabla_\theta g$, $\nabla_\theta^2 g$, $\nabla_\theta g'$, $\nabla_\theta^2 g'$ are $2r$-dominated uniformly in Θ, with $r \geq 2(2+\psi)$; (iii) $E(-\nabla_\theta^2 g(\theta))$ is negative definite, uniformly in Θ; (iv) $\omega(\cdot)$ is a bounded, twice continuously differentiable function on the interior of Γ and $\nabla_\gamma \omega(Z^t, \gamma)$ is bounded uniformly in Γ; (iv) $\nabla_\gamma \nabla_\theta g'(\theta)\omega(Z^t, \gamma)$ is continuous on $\Theta \times \Gamma$, Γ a compact subset of $\Re^d$ and is $2r$-dominated uniformly in

$\Theta \times \Gamma$, with $r \geq 2(2+\psi)$.

Assumption 4.2: (i) $E(g^{'}(y_t - \theta_1^{\dagger} y_{t-1})) < E(g^{'}(y_t - \theta_1 y_{t-1})), \forall \theta \neq \theta^{\dagger}$; (ii) $\inf_{\gamma} E(g^{'}(y_t - \theta_{2,1}^{\dagger}(\gamma) y_{t-1} + \theta_{2,2}^{\dagger}(\gamma)\omega(Z^{t-1},\gamma))) < E(g^{'}(y_t - \theta_{2,1}(\gamma) y_{t-1} + \theta_{2,2}(\gamma)\omega(Z^{t-1},\gamma))), \forall \theta \neq \theta^{\dagger}(\gamma)$.

Assumption 4.3: $T = R + n$, and as $T \to \infty$, $n/R \to \pi$, with $0 \leq \pi < \infty$.

PROPOSITION 4.1 (From Theorem 1 in [24]): With Assumptions 4.1–4.3, the following results hold: (i) Under the null,

$$M_n \xrightarrow{d} \int Z(\gamma)^2 \phi(\gamma) d\gamma$$

where Z is a Gaussian process with covariance structure,

$$\begin{aligned} K(\gamma_1, \gamma_2) = {} & S_{gg}(\gamma_1, \gamma_2) + 2\Pi \mu_{\gamma_1} A^{\dagger} S_{hh} A^{\dagger} \mu_{\gamma_2} \\ & + \Pi \mu_{\gamma_1}^{'} A^{\dagger} S_{gh}(\gamma_2) + \Pi \mu_{\gamma_2}^{'} A^{\dagger} S_{gh}(\gamma_1) \end{aligned}$$

with $\mu_{\gamma_1} = E(\nabla_{\theta_1}(g^{'}(u_t)\omega(Z^t, \gamma_1)))$, $A^{\dagger} = (-E(\nabla_{\theta_1}^2 q(u_t)))^{-1}$, and

$$S_{gg}(\gamma_1, \gamma_2) = \sum_j E(g^{'}(u_{s+1})\omega(Z^s, \gamma_1) g^{'}(u_{s+j+1})\omega(Z^{s+j}, \gamma_1))$$

$$S_{hh} = \sum_j E(\nabla_{\theta_1} q(u_s) \nabla_{\theta_1} q(u_{s+j})^{'})$$

$$S_{gh}(\gamma_1) = \sum_j E(g^{'}(u_{s+1})\omega(Z^s, \gamma_1) \nabla_{\theta_1} q(u_{s+j})^{'})$$

and γ, γ_1 and γ_2 are generic elements of Γ.
(ii) Under the alternative, for $\varepsilon > 0$ and $\delta < 1$,

$$\lim_{n \to \infty} \Pr\left(n^{-\delta} \int m_n(\gamma)^2 \phi(\gamma) d\gamma > \varepsilon \right) = 1$$

The limiting distribution under the null is a Gaussian process with a covariance structure that reflects both the time dependence and the parameter estimation error. Therefore the critical values cannot be tabulated. Valid asymptotic critical values can be constructed by using the block bootstrap for recursive estimation schemes, as detailed in [11]. In particular, define,

$$\widetilde{\theta}_{1,t}^{*} = \arg\min_{\theta_1} \frac{1}{t} \sum_{j=2}^{t} [g(y_j^{*} - \theta_1 y_{j-1}^{*}) - \theta_1^{'} \frac{1}{T} \sum_{i=2}^{T} \nabla_{\theta} g(y_i - \widehat{\theta}_1 y_{i-1})]$$

Then the bootstrap statistic is,

$$M_n^* = \int m_n^*(\gamma)^2 \phi(\gamma) d\gamma$$

where

$$m_n^*(\gamma) = n^{-1/2} \sum_{t=R}^{T-1} \left(g'(u_t^*) \omega(Z^{*,t}, \gamma) - T^{-1} \sum_{i=1}^{T-1} g'(\widehat{u}_t) \omega(Z^i, \gamma) \right)$$

Assumption 4.4: For any t, s and $\forall i, j, k = 1, 2$, and for $\Delta < \infty$,

$$\text{(i)} \quad E(\sup_{\theta, \gamma, \gamma^+} |g'(\theta) \omega(Z^{t-1}, \gamma) \nabla_\theta^k g'(\theta) \omega(Z^{s-1}, \gamma^+)|^4) < \Delta$$

where $\nabla_\theta^k(\cdot)$ denotes the k-th element of the derivative of its argument with respect to θ.

$$\text{(ii)} \quad E(\sup_\theta |\nabla_\theta^k (\nabla_\theta^i g(\theta)) \nabla_\theta^j g(\theta)|^4) < \Delta$$

and

$$\text{(iii)} \quad E(\sup_{\theta, \gamma} |g'(\theta) \omega(Z^{t-1}, \gamma) \nabla_\theta^k (\nabla_\theta^j g(\theta))|^4) < \Delta$$

PROPOSITION 4.2 (From Proposition 5 in [11]): With Assumptions 4.1–4.4, also assume that as $T \to \infty$, $l \to \infty$, and $l/T^{1/4} \to 0$, then as $T, n, R \to \infty$,

$$\Pr \left(\sup_\delta \left| \overset{*}{\Pr}(\int m_n^*(\gamma)^2 \phi(\gamma) d\gamma \leq \delta) - Pr(\int m_n(\gamma)^2 \phi(\gamma) d\gamma \leq \delta) \right| > \varepsilon \right) \to 0$$

The above proposition justifies the bootstrap procedure. For all samples except a set with probability measure approaching zero, M_n^* mimics the limiting distribution of M_n under the null, ensuring asymptotic size equal to α. Under the alternative, M_n^* still has a well defined limiting distribution, while M_n explodes, ensuring unit asymptotic power.

In closing, note that $\widetilde{\theta}_{1,t}^*$ can be replaced with $\theta_{1,t}^*$ if parameter estimation error is assumed to be asymptotically negligible. In this case, critical values are constructed via standard application of the block bootstrap.

4 Comparison of multiple models

The predictive accuracy tests that we have introduced to this point are all used to choose between two competing models. However, an even more common situation is when multiple (more than two) competing models are available, and the objective

is to assess whether there exists at least one model that outperforms a given "benchmark" model. If we sequentially compare each of the alternative models with the benchmark, we induce the so-called "data snooping" problem, where sequential test bias results in the size of our test increasing to unity, so that the null hypothesis is rejected with probability one, even when the null is true. In this subsection, we review several tests for comparing multiple models and addressing the issue of data snooping.

4.1 A reality check for data snooping

[10] proposes a test called the "reality check", which is suitable for comparing multiple models. We use the same notation as that used when discussing the DM test, except that there are now multiple alternative models, i.e. model $i = 0, 1, 2, ..., m$. Recall that $i = 0$ denotes the benchmark model. Define the following test statistic,

$$\widehat{S}_n = \max_{i=1,...,m} \widehat{S}_n(0,i) \tag{4}$$

where

$$\widehat{S}_n(0,i) = \frac{1}{\sqrt{n}} \sum_{t=R}^{T-1} (g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{i,t+1})), \quad i = 1, ..., m$$

The reality check tests the following null hypothesis:

$$H_0 : \max_{i=1,...,m} E(g(u_{0,t+1}) - g(u_{i,t+1})) \leq 0$$

against

$$H_A : \max_{i=1,...,m} E(g(u_{0,t+1}) - g(u_{i,t+1})) > 0$$

The null hypothesis states that no competing model amongst the set of m alternatives yields more accurate forecasts than the benchmark model, for a given loss function; while the alternative hypothesis states that there is at least one alternative model that outperforms the benchmark model. By jointly considering all alternative models, the reality check controls the family-wise error rate (FWER), thus circumventing the issue of data snooping, i.e. sequential test bias.

Assumption 3.1: (i) $f_i(\cdot, \theta_i^{\dagger})$ is twice continuously differentiable on the interior of Θ_i and the elements of $\nabla_{\theta_i} f_i(Z^t, \theta_i)$ and $\nabla^2_{\theta_i} f_i(Z^t, \theta_i)$ are p-dominated on Θ_i, for $i = 1, ..., m$, with $p > 2(2 + \psi)$, where ψ is the same positive constant defined in Assumption 1.1; (ii) $g(\cdot)$ is positively valued, twice continuously differentiable on Θ_i, and $g(\cdot)$, $g^{'}(\cdot)$, and $g^{''}(\cdot)$ are p-dominated on Θ_i, with p defined in (i); and (iii) let $c_{ii} = \lim_{T\to\infty} \mathrm{Var}\left(T^{-1/2} \sum_{t=1}^{T} (g(u_{0,t+1}) - g(u_{i,t+1}))\right)$, $i = 1, ..., m$, define analogous covariance terms, c_{ji}, $j, i = 1, ..., m$, and assume that c_{ji} is positive semi-definite.

PROPOSITION 3.1 (Parts (i) and (iii) are from Proposition 2.2 in [10]): With Assumptions 1.1, 1.2 and 3.1, then under the null,

$$\max_{i=1,...,m} \left(\widehat{S}_n(0,i) - \sqrt{n} E\left(g(u_{0,t+1}) - g(u_{i,t+1}) \right) \right) \xrightarrow{d} \max_{i=1,...,m} S(0,i)$$

where $S = (S(0,1),...,S(0,m))'$ is a zero mean Gaussian process with covariance matrix given by V, with V an $m \times m$ matrix, and: (i) If parameter estimation error vanishes, then for $i = 0,...,m$,

$$V = S_{g_i g_i} = \sum_{\tau=-\infty}^{\infty} E\left(g(u_{0,1}) - g(u_{i,1}) \right) \left(g(u_{0,1+\tau}) - g(u_{i,1+\tau}) \right)$$

(ii) If parameter estimation error does not vanish, then

$$\begin{aligned} V =& S_{g_i g_i} + 2\Pi \mu_0' A_0^{\dagger} C_{00} A_0^{\dagger} \mu_0 + 2\Pi \mu_i' A_i^{\dagger} C_{ii} A_i^{\dagger} \mu_i \\ & - 4\Pi \mu_0' A_0^{\dagger} C_{0i} A_i^{\dagger} \mu_i + 2\Pi S_{g_i q_0} A_0^{\dagger} \mu_0 - 2\Pi S_{g_i q_i} A_i^{\dagger} \mu_i \end{aligned}$$

where

$$C_{ii} = \sum_{\tau=-\infty}^{\infty} E(\nabla_{\theta_i} q_i(y_{1+s}, Z^s, \theta_i^{\dagger}))(\nabla_{\theta_i} q_i(y_{1+s+\tau}, Z^{s+\tau}, \theta_i^{\dagger}))'$$

$$S_{g_i q_i} = \sum_{\tau=-\infty}^{\infty} E\left((g(u_{0,1}) - g(u_{i,1})) \right) (\nabla_{\theta_i} q_i(y_{1+s+\tau}, Z^{s+\tau}, \theta_i^{\dagger}))'$$

$A_i^{\dagger} = (E(-\nabla_{\theta_i}^2 q_i(y_t, Z^{t-1}, \theta_i^{\dagger})))^{-1}$, $\mu_i = E(\nabla_{\theta_i} g(u_{i,t+1}))$, and $\Pi = 1 - \pi^{-1}\ln(1+\pi)$. (iii) Under the alternative, $\Pr(n^{-1/2}|S_n| > \varepsilon) \to 1$ as $n \to \infty$.

Of particular note is that since the maximum of a Gaussian process is not Gaussian, in general, the construction of critical values for inference is not straightforward. [10] proposes two alternatives. The first is a simulation-based approach starting from a consistent estimator of V, say $\widehat{V}$. With $\widehat{V}$, for each simulation $s = 1,...,S$, one realization is drawn from m-dimensional $N(0,\widehat{V})$ and the maximum value over $i = 1,...,m$ is recorded. Repeat this procedure for S times, with a large S, and use the $(1-\alpha)$-percentile of the empirical distribution of the maximum values. A main drawback to this approach is that we need to first estimate the covariance structure V. However, if m is large and the prediction errors exhibit a high degree of heteroskedasticity and time dependence, the estimator of V becomes imprecise and thus the inference unreliable, especially in finite samples. The second approach relies on bootstrap procedures to construct critical values, which overcomes the problem of the first approach. We resample blocks of $g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{i,t+1})$, and for each bootstrap replication $b = 1,...,B$, we calculate

$$\widehat{S}_n^{*(b)}(0,i) = n^{-1/2} \sum_{t=R}^{T-1} (g^*(\widehat{u}_{0,t+1}) - g^*(\widehat{u}_{i,t+1})) \tag{5}$$

and the bootstrap statistic is given by

$$S_n^* = \max_{i=1,\ldots,m} |\widehat{S}_n^{*(b)}(0,i) - \widehat{S}_n(0,i)|$$

the $(1-\alpha)$-percentile of the empirical distribution of B bootstrap statistics is then used for inference. Note that in [10], parameter estimation error is assumed to be asymptotically negligible. In light of this, [11] suggest a "re-centering" bootstrap procedure in order to explicitly handle the issue of non-vanishing parameter estimation error, when constructing critical values for this test. The new bootstrap statistic is defined as,

$$S_n^{**} = \max_{i=1,\ldots,m} S_n^{**}(0,i)$$

where

$$\begin{aligned} S_n^{**}(0,i) =& n^{-1/2} \sum_{t=R}^{T-1} [(g(y_{t+1}^* - f_0(Z^{*,t}, \widetilde{\theta}_{0,t}^*)) - g(y_{t+1}^* - f_i(Z^{*,t}, \widetilde{\theta}_{i,t}^*))) \\ & - \frac{1}{T} \sum_{j=1}^{T-1} (g(y_{j+1} - f_0(Z^j, \widehat{\theta}_{0,t})) - g(y_{j+1} - f_i(Z^j, \widehat{\theta}_{i,t})))] \end{aligned}$$

Note that $S_n^{**}(0,i)$ is different from the standard bootstrap statistic in Equation (5), which is defined as the difference between the statistic constructed using original samples and that using bootstrap samples. The $(1-\alpha)$-percentile of the empirical distribution of S_n^{**} can be used to construct valid critical values for inference in the case of non-vanishing parameter estimation error. Proposition 2 in [11] establishes the first order validity for the recursive estimation scheme and [12] outline the approach to constructing valid bootstrap critical values for the rolling window estimation scheme. Finally, note that [11] explain how to use the simple block bootstrap for constructing critical values when parameter estimation error is assumed to be asymptotically negligible. This procedure is perhaps the most obvious method to use for constructing critical values as it involves simply resampling the original data, carrying out the same forecasting procedures as used using the original data, and then constructing bootstrap statistics. These bootstrap statistics can be used (after subtracting the original test statistic from each of them) to form an empirical distribution which mimics the distribution of the test statistic under the null hypothesis. Finally, the empirical distribution can be used to construct critical values, which are the $(1-\alpha)$-quantiles of said distribution.

From Equation (4) and Proposition 3.1, it is immediate to see that the reality check can be rather conservative when a many alternative models are strictly dominated by the benchmark model. This is because those "bad" models do not contribute to the test statistic, simply because they are ruled out by the maximum, but contribute to the bootstrap statistics. Therefore, when many inferior models are included, the probability of rejecting the null hypothesis is actually smaller than α. Indeed, it is only for the least favorable case, in which $E(g(u_{0,t+1}) - g(u_{i,t+1})) = 0, \forall i$, that the distribution of $\widehat{S}_n$ coincides with that of

$$\max_{i=1,\ldots,m}\left(\widehat{S}_n(0,i)-\sqrt{n}E\left(g(u_{0,t+1})-g(u_{i,t+1})\right)\right)$$

We introduce two approaches for addressing the conservative nature of this test below.

4.2 A test for superior predictive ability

[13] proposes a modified reality check called the superior predictive ability (SPA) test that controls the FWER and addresses the inclusion of inferior models. The SPA test statistic is defined as,

$$T_n=\max\left\{0,\max_{i=1,\ldots,m}\frac{\widehat{S}_n(0,i)}{\sqrt{\widehat{v}_{i,i}}}\right\}$$

where $\widehat{v}_{i,i}=\frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{n}\sum_{t=R}^{T-1}((g(\widehat{u}_{0,t+1})-g(\widehat{u}_{i,t+1}))-(g(\widehat{u}^*_{0,t+1})-g(\widehat{u}^*_{i,t+1})))^2\right)$.

The bootstrap statistic is then defined as,

$$T_n^{*(b)}=\max\left\{0,\max_{i=1,\ldots,m}\{\frac{n^{-1/2}\sum_{t=R}^{T-1}(\widehat{d}^{*(b)}_{i,t}-\widehat{d}_{i,t}\mathbf{1}_{\{\widehat{d}_{i,t}\geq -A_{T,i}\}})}{\sqrt{\widehat{v}_{i,i}}}\}\right\}$$

where $\widehat{d}^{*(b)}_{i,t}=g(\widehat{u}^*_{0,t+1})-g(\widehat{u}^*_{i,t+1})$, $\widehat{d}_{i,t}=g(\widehat{u}_{0,t+1})-g(\widehat{u}_{i,t+1})$, and $A_{T,i}=\frac{1}{4}T^{-1/4}\sqrt{\widehat{v}_{i,i}}$.

The idea behind the construction of SPA bootstrap critical values is that when a competing model is too slack, the corresponding bootstrap moment condition is not re-centered, and the bootstrap statistic is not affected by this model. Therefore, the SPA test is less conservative than the reality check. [14] derive a general class of SPA tests using the generalized moment selection approach of [15] and show that Hansen's SPA test belongs to this class. [16] propose a multiple step extension of the reality check which ensures tighter control of irrelevant models.

4.3 A test based on sub-sampling

The conservative property of the reality check can be alleviated by using the sub-sampling approach to constructing critical values, at the cost of sacrificing power in finite samples. Critical values are obtained from the empirical distribution of a sequence of statistics constructed using subsamples of size $\widetilde{b}$, where $\widetilde{b}$ grows with the sample size, but at a slower rate (see [17]).

In the context of the reality check, as $n\to\infty,\widetilde{b}\to\infty$, and $\widetilde{b}/n\to 0$, define

$$S_{n,a,\widetilde{b}}=\max_{i=1,\ldots,m}S_{n,a,\widetilde{b}}(0,i),\quad a=R,\ldots,T-\widetilde{b}-1$$

where

$$S_{n,a,\widetilde{b}}(0,i) = \widetilde{b}^{-1/2} \sum_{t=a}^{a+\widetilde{b}-1} (g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{i,t+1}))$$

We obtain the empirical distribution of $T-\widetilde{b}-1$ statistics, $S_{n,a,\widetilde{b}}$, and reject the null if the test statistic $\widehat{S}_n$ is greater than the $(1-\alpha)$-quantile of the empirical distribution. The advantage of the sub-sampling approach over the bootstrap is that the test has correct size when $\max_{i=1,\ldots,m} E(g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{i,t+1})) < 0$ for some i, while the bootstrap approach delivers a conservative test in this case. However, although the sub-sampling approach ensures that the test has unit asymptotic power, the finite sample power may be rather low, since $S_{n,a,\widetilde{b}}$ diverges at rate $\sqrt{\widetilde{b}}$ instead of $\sqrt{n}$, under the alternative. Finally, note that the sub-sampling approach is also valid in the case of non-vanishing parameter estimation error because each statistic constructed using subsamples properly mimics the distribution of actual statistic.

Part II: Forecast Evaluation Using Density Based Predictive Accuracy Tests

In Part I, we introduced a variety of tests designed for comparing models based on point forecast accuracy. However, there are many practical situations in which economic decision making crucially depends not only on conditional mean forecasts (e.g. point forecasts), but also on predictive confidence intervals or predictive conditional distributions (also called predictive densities). One such case, for instance, is when value at risk (VaR) measures are used in risk management for assessment of the amount of projected financial losses due to extreme tail behavior, e.g. catastrophic events. Another common case is when economic agents are undertaking to optimize their portfolio allocations, in which case the joint distribution of multiple assets is required to be modeled and fully understood. The purpose of this section is to discuss recent tests for comparing (potentially misspecified) conditional distribution models.

5 The Kullback-Leibler information criterion approach

A well-known measure of distributional accuracy is the Kullback-Leibler Information Criterion (KLIC). Using the KLIC involves simply choosing the model which minimizes the KLIC (see, e.g., [27], [28], [29], [30]). Of note is that [27] shows that quasi maximum likelihood estimators minimize the KLIC, under mild conditions. In order to implement the KLIC, one might choose model 0 over model 1, if

$$E(\ln f_0(y_t|Z^t,\theta_0^\dagger) - \ln f_1(y_t|Z^t,\theta_1^\dagger)) > 0$$

For the i.i.d case, [28] suggests using a likelihood ratio test for choosing the conditional density model that is closer to the "true" conditional density, in terms of the KLIC. [29] suggests using a weighted version of the likelihood ratio test proposed in [28] for the case of dependent observations, while [30] employs a KLIC-based approach to select among misspecified conditional models that satisfy given moment conditions. Furthermore, the KLIC approach has recently been employed for the evaluation of dynamic stochastic general equilibrium models (see e.g., [31], [32], and [33]). For example, [32] show that the KLIC-best model is also the model with the highest posterior probability.

The KLIC is a sensible measure of accuracy, as it chooses the model which on average gives higher probability to events which have actually occurred. Also, it leads to simple likelihood ratio type tests which have a standard limiting distribution and are not affected by problems associated with accounting for parameter estimation error. However, it should be noted that if one is interested in measuring accuracy over a specific region, or in measuring accuracy for a given conditional confidence interval, say, this cannot be done in as straightforward manner using the KLIC. For example, if we want to evaluate the accuracy of different models for approximating

the probability that the rate of inflation tomorrow, given the rate of inflation today, will be between 0.5% and 1.5%, say, we can do so quite easily using the square error criterion, but not using the KLIC.

6 A predictive density accuracy test for comparing multiple misspecified models

[34] (CSa) and [35] (CSb) introduce a measure of distributional accuracy, which can be interpreted as a distributional generalization of mean square error. In addition, [34] apply this measure to the problem of selecting amongst multiple misspecified predictive density models. In this section we discuss these contributions to the literature.

Consider forming parametric conditional distributions for a scalar random variable, y_t, given Z^t, where $Z^t = (y_{t-1}, ..., y_{t-s_1}, X_t, ..., X_{t-s_{2+1}})$, with $s1, s2$ finite. With a little abuse of notation, now we define the group of conditional distribution models, from which one wishes to select a "best" model, as

$$\{F_i(u|Z^t, \theta_i^\dagger)\}_{i=1,...,m},$$

and define the true conditional distribution as

$$F_0(u|Z^t, \theta_0) = \Pr(y_{t+1} \leq u|Z^t)$$

Assume that $\theta_i^\dagger \in \Theta_i$, where Θ_i is a compact set in a finite dimensional Euclidean space, and let $\theta_i^\dagger$ be the probability limit of a quasi maximum likelihood estimator (QMLE) of the parameters of the conditional distribution under model i. If model i is correctly specified, then $\theta_i^\dagger = \theta_0$. If $m > 2$, follow [10]. Namely, choose a particular conditional distribution model as the "benchmark" and test the null hypothesis that no competing model can provide a more accurate approximation of the "true" conditional distribution, against the alternative that at least one competitor outperforms the benchmark model. Needless to say, pairwise comparison of alternative models, in which no benchmark need be specified, follows as a special case.

In this context, measure accuracy using the above distributional analog of mean square error. More precisely, define the mean square (approximation) error associated with model i, in terms of the average over U of $E\left((F_i(u|Z^t, \theta_i^\dagger) - F_0(u|Z^t, \theta_0))^2\right)$, where $u \in U$, and U is a possibly unbounded set on the real line, and the expectation is taken with respect to the conditioning variables. In particular, model 1 is more accurate than model 2, if

$$\int_U E((F_1(u|Z^t, \theta_1^\dagger) - F_0(u|Z^t, \theta_0))^2 - (F_2(u|Z^t, \theta_2^\dagger) - F_0(u|Z^t, \theta_0))^2)\phi(u)du < 0$$

where $\int_U \phi(u)du = 1$ and $\phi(u)du \geq 0$, $\forall u \in U \in \Re$.

This measure integrates over different quantiles of the conditional distribution. For any given evaluation point, this measure defines a norm and it implies a standard goodness of fit measure. Note that this measure of accuracy leads to straightforward evaluation of distributional accuracy over a given region of interest, as well as to straightforward evaluation of specific quantiles. A conditional confidence interval version of the above condition which is more natural to use in applications involving predictive interval comparison follows immediately, and can be written as

$$E\Big(((F_1(\bar{u}|Z^t,\theta_1^\dagger)-F_1(\underline{u}|Z^t,\theta_1^\dagger))-(F_1(\bar{u}|Z^t,\theta_0)-F_1(\underline{u}|Z^t,\theta_0)))^2$$

$$-((F_2(\bar{u}|Z^t,\theta_2^\dagger)-F_2(\underline{u}|Z^t,\theta_2^\dagger))-(F_1(\bar{u}|Z^t,\theta_0)-F_1(\underline{u}|Z^t,\theta_0)))^2\Big)\leq 0$$

Hereafter, $F_1(\cdot|\cdot,\theta_1^\dagger)$ is taken as the benchmark model, and the objective is to test whether some competitor model can provide a more accurate approximation of $F_0(\cdot|\cdot,\theta_0)$ than the benchmark.The null and the alternative hypotheses are:

$$H_0:\max_{i=2,\ldots,m}\int_U E((F_1(u|Z^t,\theta_1^\dagger)-F_0(u|Z^t,\theta_0))^2$$

$$-(F_i(u|Z^t,\theta_i^\dagger)-F_0(u|Z^t,\theta_0))^2)\phi(u)du\leq 0$$

versus

$$H_A:\max_{i=2,\ldots,m}\int_U E((F_1(u|Z^t,\theta_1^\dagger)-F_0(u|Z^t,\theta_0))^2$$

$$-(F_i(u|Z^t,\theta_i^\dagger)-F_0(u|Z^t,\theta_0))^2)\phi(u)du>0,$$

where $\phi(u)\geq 0$ and $\int_U\phi(u)=1$, $u\in U\in\Re$, U possibly unbounded. Note that for a given u, we compare conditional distributions in terms of their (mean square) distance from the true distribution. We then average over U. As discussed above, a possibly more natural version of the above hypotheses is in terms of conditional confidence intervals evaluation, so that the objective is to "approximate" $\Pr(\underline{u}\leq Y_{t+1}\leq\bar{u}|Z^t)$, and hence to evaluate a region of the predictive density. In that case, the null and alternative hypotheses can be stated as:

$$H_0':\max_{i=2,\ldots,m}E(((F_1(\bar{u}|Z^t\ ,\ \theta_1^\dagger)-F_1(\underline{u}|Z^t\ ,\ \theta_1^{\mathrm{f}}))$$

$$-(F_0(\bar{u}|Z^t,\ \theta_0)-F_0(\underline{u}|Z^t,\ \theta_0)))^2$$

$$-((F_i(\bar{u}|Z^t,\ \theta_i^\dagger)-F_i(\underline{u}|Z^t,\ \theta_i^\dagger))$$

$$-(F_0(\bar{u}|Z^t,\ \theta_0)-F_0(\underline{u}|Z^t,\ \theta_0)))^2)\ \leq 0$$

versus

$$H_A':\max_{i=2,\ldots,m}E(((F_1(\bar{u}|Z^t\ ,\ \theta_1^\dagger)-F_1(\underline{u}|Z^t\ ,\ \theta_1^{\mathrm{f}}))$$

$$-(F_0(\bar{u}|Z^t,\ \theta_0)-F_0(\underline{u}|Z^t,\ \theta_0)))^2$$

$$-((F_k(\overline{u}|Z^t,\ \theta_i^\dagger) - F_i(\underline{u}|Z^t,\ \theta_i^\dagger))$$
$$-(F_0(\overline{u}|Z^t,\ \theta_0) - F_0(\underline{u}|Z^t,\ \theta_0)))^2)\ > 0$$

Alternatively, if interest focuses on testing the null of equal accuracy of two conditional distribution models, say F_1 and F_i, we can simply state the hypotheses as:

$$H_0'' : \int_U E((F_1(u|Z^t,\ \theta_1^\dagger) - F_0(u|Z^t,\ \theta_0))^2$$
$$-\ (F_i\ (u|Z^t,\ \theta_i^\dagger) - F_0(u|Z^t,\ \theta_0))^2)\phi(u)\mathrm{d}u = 0$$

versus

$$H_A'' : \int_U E((F_1(u|Z^t,\ \theta_1^\dagger) - F_0(u|Z^t,\ \theta_0))^2$$
$$-\ (F_i\ (u|Z^t,\ \theta_i^\dagger) - F_0(u|Z^t,\ \theta_0))^2)\phi(u)\mathrm{d}u \neq 0,$$

or we can write the predictive density (interval) version of these hypotheses.

Of course, we do not know $F_0(u|Z^t)$. However, it is easy to see that

$$\begin{aligned} E((F_1(u|Z^t,\ \theta_1^\dagger) - F_0(u|Z^t,\ \theta_0))^2 - (F_i(u|Z^t,\ \theta_i^\dagger) - F_0(u|Z^t,\ \theta_0))^2) \\ = E((1\{y_{t+1}\ \leq u\} - F_1(u|Z^t,\ \theta_1^\dagger))^2) \\ - E\ ((1\{y_{t+1}\ \leq u\} - F_i\ (u|Z^t,\ \theta_i^\dagger))^2), \end{aligned} \tag{6}$$

where the right-hand side of Equation (6) does not require any knowledge of the true conditional distribution.

The intuition behind Equation (6) is very simple. First, note that for any given u, $E(1\{y_{t+1}\ \leq\ u\}|Z^t) = \Pr(y_{t+1}\ \leq\ u|Z^t) = F_0(u|Z^t,\ \theta_0)$. Thus, $1\{y_{t+1}\ \leq\ u\} - F_i(u|Z^t,\ \theta_i^\dagger)$ can be interpreted as an "error" term associated with computation of the conditional expectation under F_i. Now, for $i = 1, ..., m$:

$$\mu_i^2(u) = E(\ (1\{y_{t+1}\ \leq u\} - F_i\ (u|Z^t,\ \theta_i^\dagger))^2)$$
$$= E(((1\{y_{t+1}\ \leq u\} - F_0(u|Z^t,\ \theta_0)) - (F_i(u|Z^t,\ \theta_i^\dagger) - F_0(u|Z^t,\ \theta_0)))^2)$$
$$= E((1\{y_{t+1}\ \leq u\} - F_0(u|Z^t,\ \theta_0))^2) + E\ ((F_i\ (u|Z^t,\ \theta_i^\dagger) - F_0(u|Z^t,\ \theta_0))^2),$$

given that the expectation of the cross product is zero (which follows because $1\ \{y_{t+1} \leq u\} - F_0(u|Z^t,\ \theta_0)$ is uncorrelated with any measurable function of Z^t). Therefore,

$$\begin{aligned} \mu_1^2(u) - \mu_i^2(u) = E(\ (F_1\ (u|Z^t,\ \theta_1^\dagger) - F_0(u|Z^t,\ \theta_0))^2) \\ - E\ ((F_i\ (u|Z^t,\ \theta_i^\dagger) - F_0(u|Z^t,\ \theta_0))^2) \end{aligned} \tag{7}$$

The statistic of interest is

$$Z_{n,j} = \max_{i=2,...m} \int_U Z_{n,u,j}(1,\ i)\phi(u)\mathrm{d}u, \quad j = 1, 2,$$

where for $j = 1$ (rolling estimation scheme),

$$Z_{n,u,1}(1,\ i) = \frac{1}{\sqrt{n}} \sum_{t=R}^{T-1} ((1\{y_{t+1}\ \leq u\} - F_1(u|Z^t,\ \widehat{\theta}_{1,t,\mathrm{rol}}))^2$$

$$- (1\{y_{t+1}\ \leq u\} - F_i(u|Z^t,\ \widehat{\theta}_{i,t,\mathrm{rol}}))^2)$$

and for $j = 2$ (recursive estimation scheme),

$$Z_{n,u,2}(1,\ i) = \frac{1}{\sqrt{n}} \sum_{\mathrm{t}=R}^{T-1} ((1\{y_{t+1}\ \leq u\} - F_1(u|Z^t,\ \widehat{\theta}_{1,t,\mathrm{rec}}))^2$$

$$- (1\{y_{t+1}\ \leq u\} - F_i(u|Z^t,\ \widehat{\theta}_{i,t,\mathrm{rec}}))^2),$$

where $\widehat{\theta}_{i,t,\mathrm{rol}}$ and $\widehat{\theta}_{i,t,\mathrm{rec}}$ are defined as:

$$\widehat{\theta}_{i,t,\mathrm{rol}} = \arg\min_{\theta\in\Theta} \frac{1}{R} \sum_{j=t-R+1}^{t} q(y_j, Z^{j-1}, \theta),\ R \leq t \leq T-1$$

and

$$\widehat{\theta}_{i,t,\mathrm{rec}} = \arg\min_{\theta\in\Theta} \frac{1}{t} \sum_{j=1}^{t} q(y_j, Z^{j-1}, \theta),\ t = R, R+1, R+n-1$$

As shown above and in [34], the hypotheses of interest can be restated as:

$$H_0 : \max_{i=2,\ldots,m} \int_U (\mu_1^2(u) - \mu_i^2(u))\phi(u)\mathrm{d}u \leq 0$$

versus

$$H_A : \max_{i=2,\ldots,m} \int_U (\mu_1^2(u) - \mu_i^2(u))\phi(u)\mathrm{d}u > 0$$

where $\mu_i^2(u) = E(\ (1\{y_{t+1}\ \leq u\} - F_i\ (u|Z^t,\ \theta_i^\dagger))^2)$.

Assumption 6.1: (i) $\theta_i^\dagger$ is uniquely defined,

$$E(\ln(f_i(y_t, Z^{t-1}, \theta_i))) < E(\ln(f_i(y_t, Z^{t-1}, \theta_i^\dagger))),$$

for any $\theta_i \neq \theta_i^\dagger$; (ii) $\ln f_i$ is twice continuously differentiable on the interior of Θ_i, and $\forall\Theta_i$ a compact subset of $\Re^{p(i)}$; (iii) the elements of $\nabla_{\theta_i} \ln f_i$ and $\nabla^2_{\theta_i} \ln f_i$ are p-dominated on Θ_i, with $p > 2(2+\psi)$, where ψ is the same positive constant as defined in Assumption 1.1; and (iv) $E(-\nabla^2_{\theta_i} \ln f_i)$ is negatively definite uniformly on Θ_i.

Assumption 6.2: $T = R+n$, and as $T \to \infty$, $n/R \to \pi$, with $0 < \pi < \infty$.

Assumption 6.3: (i) $F_i(u|Z^t,\theta_i)$ is continuously differentiable on the interior of Θ_i and $\nabla_{\theta_i}F_i(u|Z^t,\theta_i^\dagger)$ is $2r$-dominated on Θ_i, uniformly in u, $r>2$, $\forall i$;[1] and (ii) let

$$v_{ii}(u)=\text{plim}_{T\to\infty}\text{Var}\Big(\frac{1}{\sqrt{T}}\sum_{t=s}^{T}(((1\{y_{t+1}\le u\}-F_1(u|Z^t,\theta_1^\dagger))^2-\mu_1^2(u))$$

$$-((1\{y_{t+1}\le u\}-F_i(u|Z^t,\theta_i^\dagger))^2-\mu_i^2(u))\Big),\ \forall i$$

define analogous covariance terms, $v_{j,i}(u), j,i=2,...,m$, and assume that $[v_{j,i}(u)]$ is positive semi-definite, uniformly in u.

PROPOSITION 6.1 (From Proposition 1 in [35]): With Assumptions 1.1, 6.1–6.3, then

$$\max_{i=2,...,m}\int_U(Z_{n,u,j}\ (1\ ,\ i)-\sqrt{n}(\mu_1^2(u)-\mu_i^2(u)))\phi_U(u)\mathrm{d}u$$

$$\xrightarrow{d}\max_{i=2,...,m}\int_U Z_{1,i,j}(u)\phi_U(u)\mathrm{d}u$$

where $Z_{1,i,j}(u)$ is a zero mean Gaussian process with covariance $C_{i,j}(u,\ u')(j=1$ corresponds to rolling and $j=2$ to recursive estimation schemes), equal to:

$$E(\sum_{j=-\infty}^{\infty}((1\{y_{s+1}\ \le u\}-F_1(u|Z^s,\ \theta_1^\dagger))^2-\mu_1^2(u))\times\ ((1\{y_{s+j+1}\ \le u'\}$$

$$-F_1(u'|Z^{s+j},\ \theta_1^\dagger))^2-\mu_1^2(u')))+E(\sum_{j=-\infty}^{\infty}((1\{y_{s+1}\ \le u\}-F_i(u|Z^s,\ \theta_i^\dagger))^2-\mu_i^2(u))$$

$$\times\ ((1\{y_{s+j+1}\ \le u'\}-F_i(u'|Z^{s+j},\ \theta_i^\dagger))^2-\mu_i^2(u')))-2E(\sum_{j=-\infty}^{\infty}((1\{y_{s+1}\ \le u\}$$

$$-F_1(u|Z^s,\ \theta_1^\dagger))^2-\mu_1^2(u))\times\ ((1\{y_{s+j+1}\ \le u'\}-F_i(u'|Z^{s+j},\ \theta_i^\dagger))^2-\mu_i^2(u')))$$

$$+4\Pi_j m_{\theta_1^\dagger}(u)'A(\theta_1^\dagger)\times E(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(y_{s+1}|Z^s,\ \theta_1^\dagger)\nabla_{\theta_1}\ln f_1(y_{s+j+1}|Z^{s+j},\ \theta_1^\dagger)')$$

$$\times A(\theta_1^\dagger)m_{\theta_1^\dagger}(u')+4\Pi_j m_{\theta_i^\dagger}(u)'A(\theta_i^\dagger)\times E(\sum_{j=-\infty}^{\infty}\nabla_{\theta_i}\ln f_i(y_{s+1}|Z^s,\ \theta_i^\dagger)$$

$$\times\nabla_{\theta_i}\ln f_i(y_{s+j+1}|Z^{s+j},\ \theta_i^\dagger)')\times A(\theta_i^\dagger)m_{\theta_i^\dagger}(u')-4\Pi_j m_{\theta_1^\dagger}(u,\ )'A(\theta_1^\dagger)$$

$$\times E(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(y_{s+1}|Z^s,\ \theta_1^\dagger)\nabla_{\theta_i}\ln f_i(y_{s+j+1}|Z^{s+j}\times A(\theta_i^\dagger)m_{\theta_i^\dagger}(u')$$

[1] We require that for $j=1,...,p_i$, $E(\nabla_\theta F_i(u|Z^t,\theta_i^\dagger))_j\ge D_t(u)$, with $\sup_t\sup_{u\in\Re}E(D_t(u)^{2r})<\infty$.

$$-4C\Pi_j m_{\theta_1^\dagger}(u)' A(\theta_1^\dagger) \times E(\sum_{j=-\infty}^{\infty} \nabla_{\theta_1} \ln f_1(y_{s+1}|Z^s,\ \theta_1^\dagger) \times ((1\{y_{s+j+1}\ \leq u\}$$

$$-F_1\ (u|Z^{s+j},\ \theta_1^\dagger))^2 - \mu_1^2(u))) + 4C\Pi_j m_{\theta_1^\dagger}(u)' A(\theta_1^\dagger) \times E(\sum_{j=-\infty}^{\infty} \nabla_{\theta_1} \ln f_1(y_{s+1}|Z^s,\ \theta_1^\dagger)$$

$$\times((1\{y_{s+j+1}\ \leq u\} - F_i(u|Z^{s+j},\ \theta_i^\dagger))^2 - \mu_i^2(u))) - 4C\Pi_j m_{\theta_i^\dagger}(u)' A(\theta_i^\dagger)$$

$$\times E(\sum_{j=-\infty}^{\infty} \nabla_{\theta_i} \ln f_i(y_{s+1}|Z^s, \theta_i^\dagger)' \times ((1\{y_{s+j+1}\ \leq u\} - F_i(u|Z^{s+j},\ \theta_i^\dagger))^2 - \mu_i^2(u)))$$

$$+4C\Pi_j m_{\theta_i^\dagger}(u)' A(\theta_i^\dagger) \times E(\sum_{j=-\infty}^{\infty} \nabla_{\theta_i} \ln f_i(y_{s+1}|Z^s,\ \theta_i^\dagger)' \times\ ((1\{y_{s+j+1}\ \leq u\}$$

$$-F_1(u|Z^{s+j},\ \theta_1^\dagger))^2 - \mu_1^2(u)))$$

with

$$m_{\theta_i^\dagger}(u)' = E(\nabla_{\theta_i} F_i(u|Z^t,\ \theta_i^\dagger)'(1\{y_{t+1}\ \leq u\} - F_i(u|Z^t,\ \theta_i^\dagger)))$$

and

$$A(\theta_i^\dagger)\ = A_i^\dagger =\ (E(-\nabla^2_{\theta_i} \ln f_i(y_{t+1}|Z^t,\ \theta_i^\dagger)))^{-1}$$

and for $j = 1$ and $n \leq R$, $\Pi_1 = (\pi - \frac{\pi^2}{3})$, $C\Pi_1 = \frac{\pi}{2}$, and for $n > R$, $\Pi_1 = (1 - \frac{1}{3\pi})$ and $C\Pi_1 = (1 - \frac{1}{2\pi})$. Finally, for $j = 2$, $\Pi_2 = 2(1 - \pi^{-1}\ln(1+\pi))$ and $C\Pi_2 = 0.5\Pi_2$.

From this proposition, note that when all competing models provide an approximation to the true conditional distribution that is as (mean square) accurate as that provided by the benchmark (i.e. when $\int_U (\mu_1^2(u) - \mu_i^2(u))\phi(u)\mathrm{d}u = 0, \forall i$), then the limiting distribution is a zero mean Gaussian process with a covariance kernel which is not nuisance parameter free. Additionally, when all competitor models are worse than the benchmark, the statistic diverges to minus infinity at rate $\sqrt{n}$. Finally, when only some competitor models are worse than the benchmark, the limiting distribution provides a conservative test, as Z_P will always be smaller than $\max_{i=2,\ldots,m} \int_U (Z_{n,u}\ (1,\ i) - \sqrt{n}(\mu_1^2(u) - \mu_i^2(u)))\phi(u)\mathrm{d}u$, asymptotically. Of course, when H_A holds, the statistic diverges to plus infinity at rate $\sqrt{n}$.

For the case of evaluation of multiple conditional confidence intervals, consider the statistic:

$$V_{n,\tau} = \max_{i=2,\ldots,m} V_{n,\underline{u},\overline{u},\tau}(1,\ i)$$

where

$$V_{n,\underline{u},\overline{u},\tau}(1,\ i) =\ \frac{1}{\sqrt{n}} \sum_{t=R}^{T-1} ((1\{\underline{u} \leq y_{t+1}\ \leq \overline{u}\} - (F_1(\overline{u}|Z^t,\ \widehat{\theta}_{1,t,\tau})$$

$$-F_1(\underline{u}|Z^t,\ \widehat{\theta}_{1,t,\tau})))^2 - (1\{\underline{u} \leq y_{t+1}\ \leq \overline{u}\} - (F_i(\overline{u}|Z^t,\ \widehat{\theta}_{i,t,\tau}) - F_i(\underline{u}|Z^t,\ \widehat{\theta}_{i,t,\tau})))^2)$$

where $s = \max\{s1,\ s2\}$, $\tau = 1, 2$, and $\widehat{\theta}_{i,t,\tau} = \widehat{\theta}_{i,t,\mathrm{rol}}$ for $\tau = 1$, and $\widehat{\theta}_{i,t,\tau} = \widehat{\theta}_{k,t,\mathrm{rec}}$ for $\tau = 2$.

We then have the following result,

PROPOSITION 6.2 (From Proposition lb in [35]): With Assumptions 1.1, 6.1–6.3, then for $\tau = 1$,

$$\max_{i=2,...m}(V_{n,\underline{u},\overline{u},\tau}\ (1\ ,\ i)\ -\ \sqrt{n}(\mu_1^2-\mu_i^2))\ \xrightarrow{d}\ \max_{i=2,...m} V_{n,i,\tau}(\underline{u},\ \overline{u})$$

where $V_{n,i,\tau}(\underline{u},\ \overline{u})$ is a zero mean normal random variable with covariance $c_{ii} = v_{ii} + p_{ii} + cp_{ii}$, where v_{ii} denotes the component of the long-run variance matrix we would have in absence of parameter estimation error, p_{ii} denotes the contribution of parameter estimation error and cp_{ii} denotes the covariance across the two components. In particular:

$$v_{ii} = E\sum_{j=-\infty}^{\infty}(((1\{\underline{u}\le y_{s+1}\ \le\overline{u}\}-(F_1(\overline{u}|Z^s,\ \theta_1^\dagger)-F_1(\underline{u}|Z^s,\ \theta_1^\dagger)))^2-\mu_1^2)$$

$$\times((1\{\underline{u}\le y_{s+1+j}\ \le\overline{u}\}-(F_1(\overline{u}|Z^{s+j},\ \theta_1^\dagger)-F_1(\underline{u}|Z^{s+j},\ \theta_1^\dagger)))^2-\mu_1^2))$$

$$+E\sum_{j=-\infty}^{\infty}(((1\{\underline{u}\le y_{s+1}\ \le\overline{u}\}-(F_i(\overline{u}|Z^s,\ \theta_i^\dagger)-F_i(\underline{u}|Z^s,\ \theta_i^\dagger)))^2-\mu_i^2)$$

$$\times((1\{\underline{u}\le y_{s+1+j}\ \le\overline{u}\}-(F_i(\overline{u}|Z^{s+j},\ \theta_i^\dagger)-F_i(\underline{u}|Z^{s+j},\ \theta_i^\dagger)))^2-\mu_i^2))$$

$$-2E\sum_{j=-\infty}^{\infty}(((1\{\underline{u}\le y_{s+1}\ \le\overline{u}\}-(F_1(\overline{u}|Z^s,\ \theta_1^\dagger)-F_1(\underline{u}|Z^s,\ \theta_1^\dagger)))^2-\mu_1^2)$$

$$\times((1\{\underline{u}\le y_{s+1+j}\ \le\overline{u}\}-(F_i(\overline{u}|Z^{s+j},\ \theta_i^\dagger)-F_i(\underline{u}|Z^{s+j},\theta_i^\dagger)))^2-\mu_i^2))$$

Also,

$$p_{ii} = 4m'_{\theta_1^\dagger}A(\theta_1^\dagger)E(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_i(y_{s+1}|Z^s,\ \theta_1^\dagger)\nabla_{\theta_1}\ln f_i(y_{s+1+j}|Z^{s+j},\ \theta_1^\dagger)')\times A(\theta_1^\dagger)m_{\theta_1^\dagger}$$

$$+4m'_{\theta_i^\dagger}A(\theta_i^\dagger)E(\sum_{j=-\infty}^{\infty}\nabla_{\theta_i}\ln f_i(y_{s+1}|Z^s,\ \theta_i^\dagger)\nabla_{\theta_i}\ln f_i(y_{s+1+j}|Z^{s+j},\ \theta_i^\dagger)')\times A(\theta_i^\dagger)m_{\theta_i^\dagger}$$

$$-8m'_{\theta_1^\dagger}A(\theta_1^\dagger)E(\nabla_{\theta_1}\ln f_1(y_{s+1}|Z^s,\ \theta_1^\dagger)\nabla_{\theta_i}\ln f_i(y_{s+1+j}|Z^{s+j},\ \theta_i^\dagger)')\times A(\theta_i^\dagger)m_{\theta_i^\dagger}$$

Finally,

$$cp_{ii} = -4m'_{\theta_1^\dagger}A(\theta_1^\dagger)E(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(y_{s+1}|Z^s,\ \theta_1^\dagger)$$

$$\times((1\{\underline{u}\le y_{s+j}\ \le\overline{u}\}-(F_1(\overline{u}|Z^{s+j},\ \theta_1^\dagger)-F_1(\underline{u}|Z^{s+j},\ \theta_1^\dagger)))^2-\mu_1^2)$$

$$+8m'_{\theta_1^\dagger}A(\theta_1^\dagger)E(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(y_s|Z^s,\ \theta_1^\dagger)$$

$$\times ((1\{\underline{u} \leq y_{s+1+j} \leq \overline{u}\} - (F_i(\overline{u}|Z^{s+j}, \theta_i^\dagger) - F_i(\underline{u}|Z^s, \theta_i)))^2 - \mu_i^2))$$

$$-4m'_{\theta_i^\dagger} A(\theta_i^\dagger) E(\sum_{j=-\infty}^{\infty} \nabla_{\theta_i} \ln f_i(y_{s+1}|Z^s, \theta_i^\dagger)$$

$$\times ((1\{\underline{u} \leq y_{s+j} \leq \overline{u}\} - (F_i(\overline{u}|Z^{s+j}, \theta_i^\dagger) - F_i(\underline{u}|Z^{s+j}, \theta_i^\dagger)))^2 - \mu_i^2))$$

with

$$m'_{\theta_i^\dagger} = E(\nabla_{\theta_i}(F_i(\overline{u}|Z^t, \theta_i^\dagger) - F_i(\overline{u}|Z^t, \theta_i^\dagger))$$

$$\times (1\{\underline{u} \leq y_t \leq \overline{u}\} - (F_i(\overline{u}|Z^t, \theta_i^\dagger) - F_i(\overline{u}|Z^t, \theta_i^\dagger))))$$

and

$$A(\theta_i^\dagger) = (E(-\ln \nabla^2_{\theta_i} f_i(y_t|Z^t, \theta_i^\dagger)))^{-1}$$

An analogous result holds for the case where $\tau = 2$, and is omitted for the sake of brevity.

Due to the contribution of parameter estimation error, simulation error, and the time series dynamics to the covariance kernel (see Proposition 6.1), critical values cannot be directly tabulated. As a result, block bootstrap techniques are used to construct valid critical values for statistical inference. In order to show the first order validity of the bootstrap, the authors derive the limiting distribution of appropriately formed bootstrap statistics and show that they coincide with the limiting distribution given in Proposition 6.1. Recalling that as all candidate models are potentially misspecified under both hypotheses, the parametric bootstrap is not generally applicable in our context. Instead, we must begin by resampling b blocks of length $l, bl = T - 1$. Let $\mathrm{Y}_t^* = (\Delta \log X_t^*, \Delta \log X_{t-1}^*)$ be the resampled series, such that Y_2^*,..., Y_{l+1}^*, Y_{l+2}^*,..., Y_{T-l+2}^*,..., Y_T^* equals Y_{I_1+1},..., Y_{I_1+l}, Y_{I_2+1},..., Y_{I_b+1},..., Y_{I_b+T}, where I_j, $i = 1$,..., b are independent, discrete uniform random variates on 1,..., $T-1+1$. That is, $I_j = i, i = 1$,..., $T-l$ with probability $1/(T-l)$. Then, use Y_t^* to compute $\widehat{\theta}_{j,T}^*$ and plug in $\widehat{\theta}_{j,T}^*$ in order to simulate a sample under model j, $j = 1$,..., m. Let $\mathrm{Y}_{j,n}(\widehat{\theta}_{j,T}^*)$, $n = 2$,..., S denote the series simulated in this manner. At this point, we need to distinguish between the case where $\delta = 0$ (vanishing simulation error) and $\delta > 0$ (non-vanishing simulation error). In the former case, we do not need to resample the simulated series, as there is no need to mimic the contribution of simulation error to the covariance kernel. On the other hand, in the latter case we draw $\widetilde{b}$ blocks of length $\widetilde{l}$ with $\widetilde{bl} = S-1$, and let $\mathrm{Y}_{j,n}^*(\widehat{\theta}_{j,T}^*)$, $j = 1$,..., m, $n = 2$,..., S denote the resampled series under model j. Notice that $Y_{j,2}^*(\widehat{\theta}_{j,T}^*)$,..., $\mathrm{Y}_{j,l+1}^*(\widehat{\theta}_{j,T}^*)$,..., $\mathrm{Y}_{j,S}^*(\widehat{\theta}_{j,T}^*)$ is equal to $Y_{j,\widetilde{I}_1}(\widehat{\theta}_{j,T}^*)$,..., $Y_{j,\widetilde{I}_1+l}(\widehat{\theta}_{j,T}^*)$ ..., $Y_{j,\widetilde{I}\widetilde{b}_1+l}(\widehat{\theta}_{j,T}^*)$ where $\widetilde{I}_i, i = 1$,..., $\widetilde{b}$ are independent discrete uniform random variates on 1,..., $S - \widetilde{l}$. Also, note that for each of the m models, and for each bootstrap replication, we draw $\widetilde{b}$ discrete uniform random variates (the $\widetilde{I}_i$) on $1, \ldots, S - \widetilde{l}$, and that draws are independent across models. Thus, in our use of notation, we have suppressed the dependence of $\widetilde{I}_i$ on j.

Thereafter, form bootstrap statistics as follows:

$$Z^*_{n,\tau} = \max_{i=2,...m} \int_U Z^*_{n,u,\tau}(1,\ i)\phi(u)\mathrm{d}u,$$

where for $\tau = 1$ (rolling estimation scheme), and for $\tau = 2$ (recursive estimation scheme):

$$Z^*_{n,u,\tau}(1,\ i) = \frac{1}{\sqrt{n}} \sum_{t=R}^{T-1} (((1\{y^*_{t+1} \leq u\} - F_1(u|Z^{*,t}\widetilde{\theta}^*_{1,t,\tau}))^2$$

$$- (1\{y^*_{t+1} \leq u\} - F_i(u|Z^{*,t}\widetilde{\theta}^*_{i,t,\tau}))^2)$$

$$-\frac{1}{T} \sum_{j=s+1}^{T-1} ((1\{y_{j+1} \leq u\} - F_1(u|Z^i,\ \widehat{\theta}_{1,t,\tau}))^2 - (1\{y_{j+1} \leq u\} - F_i(u|Z^j,\ \widehat{\theta}_{i,t,\tau}))^2))$$

Note that each bootstrap term, say $1\{y^*_{t+1} \leq u\} - F_i(u|Z^{*,t},\ \widetilde{\theta}^*_{i,t,\tau})$, $t \geq R$, is re-centered around the (full) sample mean $\frac{1}{T}\sum_{j=s+1}^{T-1}(1\{y_{j+1} \leq u\} - F_i(u|Z^j,\ \widehat{\theta}_{i,t,\tau}))^2$. This is necessary as the bootstrap statistic is constructed using the last n resampled observations, which in turn have been resampled from the full sample. In particular, this is necessary regardless of the ratio n/R. If $n/R \to 0$, then we do not need to mimic parameter estimation error, and so could simply use $\widehat{\theta}_{1,t,\tau}$ instead of $\widetilde{\theta}^*_{1,t,\tau}$, but we still need to recenter any bootstrap term around the (full) sample mean.

Note that re-centering is necessary, even for first order validity of the bootstrap, in the case of over-identified generalized method of moments (GMM) estimators [see, e.g., [36], [37], [38], [39]]. This is due to the fact that, in the over-identified case, the bootstrap moment conditions are not equal to zero, even if the population moment conditions are. However, in the context of m-estimators using the full sample, re-centering is needed only for higher order asymptotics, but not for first order validity, in the sense that the bias term is of smaller order than $T^{-1/2}$. Namely, in the case of recursive m-estimators the bias term is instead of order $T^{-1/2}$ and so it does contribute to the limiting distribution. This points to a need for re-centering when using recursive estimation schemes.

For the confidence interval case, define:

$$V^*_{n,\tau} = \max_{i=2,...m}, V^*_{n_{\underline{u},\overline{u}},\tau}(1,\ i)$$

and

$$V^*_{n,\underline{u},\overline{u},\tau}(1,\ i) = \frac{1}{\sqrt{n}} \sum_{t=R}^{T-1} ((1\{\underline{u} \leq y^*_{t+1} \leq \overline{u}\} - (F_1(\overline{u}|Z^{*t},\widetilde{\theta}^*_{1,t,\tau}) - F_1(\underline{u}|Z^{*t},\widetilde{\theta}^*_{1,t,\tau})))^2$$

$$- (1\{\underline{u} \leq y^*_{t+1} \leq \overline{u}\} - (F_i(\overline{u}|Z^{*t},\widetilde{\theta}^*_{i,t,\tau}) - F_1(\underline{u}|Z^{*t},\widetilde{\theta}^*_{i,t,\tau})))^2)$$

$$-\frac{1}{T} \sum_{j=s+1}^{T-1} ((1\{\underline{u} \leq y_{i+1} \leq \overline{u}\} - (F_1(\overline{u}|Z^j,\ \widehat{\theta}_{1,t,\tau}) - F_1(\underline{u}|Z^j,\ \widehat{\theta}_{1,t,\tau})))^2$$

$$- (1\{\underline{u} \leq y_{j+1} \leq \overline{u}\} - (F_i(\overline{u}|Z^j, \widehat{\theta}_{i,t,\mathrm{r}}) - F_1(\underline{u}|Z^j, \widehat{\theta}_{i,t,\tau})))^2)$$

where, as usual, $\tau = 1, 2$. The following results then hold,

PROPOSITION 6.3 (From Proposition 6 in [35]): With Assumptions 1.1, 6.1–6.3, also, assume that as $T \to \infty, l \to \infty$, and that $\frac{l}{T^{1/4}} \to 0$. Then, as T, n and $R \to \infty$, for $\tau = 1, 2$:

$$\Pr\Big(\sup_{v \in \Re} |\overset{*}{\Pr_T}(\max_{i=2,\ldots m} \int_U Z^*_{n,u,\tau}(1\ ,\ i)\phi(u)\mathrm{d}u \leq v)$$

$$- \Pr\big(\max_{i=2,\ldots,m} \int_U Z^{\mu}_{n,u,\tau}\ (1\ ,\ i)\phi(u)\mathrm{d}u \leq v)| > \varepsilon\big)\ \to 0,$$

where $Z^{\mu}_{n,u,\tau}(1,\ i) = Z_{n,u,\tau}(1,\ i) - \sqrt{n}(\mu_1^2(u) - \mu_i^2(u))$, and where $\mu_1^2(u) - \mu_i^2(u)$ is defined as in Equation (7).

PROPOSITION 6.4 (From Proposition 7 in [35]): With Assumptions 1.1, 6.1–6.3, also assume that as $T \to \infty, l \to \infty$, and that $\frac{l}{T^{1/4}} \to 0$. Then, as T, n and $R \to \infty$, for $\tau = 1, 2$:

$$\Pr\Big(\sup_{v \in \Re} |\overset{*}{\Pr_T}(\max_{i=2,\ldots m}, V^*_{n,\underline{u},\overline{u},\tau}(1,\ i)\ \leq v)$$

$$- \Pr\big(\max_{i=2,\ldots m}, V^{\mu}_{n,\underline{u},\overline{u},\tau}\ (1,\ i) \leq v)|\ > \varepsilon\big)\ \to 0$$

where $V^{\mu}_{n,\underline{u},\overline{u},\tau}(1\ ,\ i) = V_{n,\underline{u},\overline{u},\tau}(1\ ,\ i) - \sqrt{n}(\mu_1^2(u) - \mu_i^2(u))$.

The above results suggest proceeding in the following manner. For brevity, consider the case of $Z^*_{n,\tau}$. For any bootstrap replication, compute the bootstrap statistic, $Z^*_{n,\tau}$. Perform B bootstrap replications (B large) and compute the quantiles of the empirical distribution of the B bootstrap statistics. Reject H_0, if $Z_{n,\tau}$ is greater than the $(1\ -\alpha)$th-percentile. Otherwise, do not reject. Now, for all samples except a set with probability measure approaching zero, $Z_{n,\tau}$ has the same limiting distribution as the corresponding bootstrapped statistic when $E(\mu_1^2(u)\ - \mu_i^2(u)) = 0, \forall i$, ensuring asymptotic size equal to α. On the other hand, when one or more competitor models are strictly dominated by the benchmark, the rule provides a test with asymptotic size between 0 and α. Under the alternative, $Z_{n,\tau}$ diverges to (plus) infinity, while the corresponding bootstrap statistic has a well defined limiting distribution, ensuring unit asymptotic power.

From the above discussion, we see that the bootstrap distribution provides correct asymptotic critical values only for the least favorable case under the null hypothesis; that is, when all competitor models are as good as the benchmark model. When $\max_{i=2,\ldots,m} \int_U (\mu_1^2(u) - \mu_i^2(u))\phi(u)\mathrm{d}u = 0$, but $\int_U (\mu_1^2(u) - \mu_i^2(u))\phi(u)\mathrm{d}u < 0$ for some i, then the bootstrap critical values lead to conservative inference. An alternative to our bootstrap critical values in this case is the construction of critical values based on subsampling, which is briefly discussed in Section 4.3. Heuristically, construct $T - 2b_T$ statistics using subsamples of length b_T , where $b_T/T \to 0$. The empirical distribution of these statistics computed over the various subsamples

properly mimics the distribution of the statistic. Thus, subsampling provides valid critical values even for the case where $\max_{i=2,\ldots,m} \int_U (\mu_1^2(u) - \mu_i^2(u))\phi(u)\mathrm{d}u = 0$, but $\int_U (\mu_1^2(u) - \mu_i^2(u))\phi(u)\mathrm{d}u < 0$ for some i. This is the approach used by [40], for example, in the context of testing for stochastic dominance. Needless to say, one problem with subsampling is that unless the sample is very large, the empirical distribution of the subsampled statistics may yield a poor approximation of the limiting distribution of the statistic. Another alternative approach for addressing the conservative nature of our bootstrap critical values is the Hansen's SPA approach (see Section 4.2 and [13]). Hansen's idea is to recenter the bootstrap statistics using the sample mean, whenever the latter is larger than (minus) a bound of order $\sqrt{2T \log\log T}$. Otherwise, do not recenter the bootstrap statistics. In the current context, his approach leads to correctly sized inference when $\max_{i=2,\ldots,m} \int_U (\mu_1^2(u) - \mu_i^2(u))\phi(u)\mathrm{d}u = 0$, but $\int_U (\mu_1^2(u) - \mu_i^2(u))\phi(u)\mathrm{d}u < 0$ for some i. Additionally, his approach has the feature that if all models are characterized by a sample mean below the bound, the null is "accepted" and no bootstrap statistic is constructed.

Part III: Forecast Evaluation Using Density Based Predictive Accuracy Tests That Are Not Loss Function Dependent: The Case of Stochastic Dominance

All predictive accuracy tests outlined in previous two parts of this chapter are loss functions dependent, i.e. loss functions such as mean squared forecast error (MSFE) and mean absolute forecast error (MAFE) must be specified prior to test construction. Evidently, given possible misspecification, model rankings may change under different loss functions. In the following section, we introduce a novel criterion for forecast evaluation that utilizes the entire distribution of forecast errors, is robust to the choice of loss function, and ranks distributions of forecast errors via stochastic dominance type tests.

7 Robust forecast comparison

[41] (JCS) introduce the concepts of general-loss (GL) forecast superiority and convex-loss (CL) forecast superiority and develop tests for GL (CL) superiority that are based on an out-of-sample generalization of the tests introduced by [42]. The JCS tests evaluate the entire forecast error distribution and do not require knowledge or specification of a loss function, i.e. tests are robust to the choice of loss function. In addition, parameter estimation error and data dependence are taken into account, and heterogeneity that is induced by distributional change over time is allowed for.

The concepts of general-loss (GL) forecast superiority and convex-loss (CL) forecast superiority are defined as follow:

(1) For any two sequences of forecast errors $u_{1,t}$ and $u_{2,t}$, $u_{1,t}$ general-loss (GL) outperforms $u_{2,t}$, denoted as $u_1 \succeq_G u_2$, if and only if $E(g(u_{1,t})) \leq E(g(u_{2,t}))$, $\forall g(\cdot) \in GL(\cdot)$, where $GL(\cdot)$ are the set of general loss functions with properties specified in [43]; and

(2) $u_{1,t}$ convex-loss (CL) outperforms $u_{2,t}$, denoted as $u_1 \succeq_C u_2$, if and only if $E(g(u_{1,t})) \leq E(g(u_{2,t}))$, $\forall g(\cdot) \in CL(\cdot)$, where $CL(\cdot)$ are the set of general loss functions which in addition are convex.

These authors also establish linkages between GL(CL) forecast superiority and first(second) order stochastic dominance, allowing for the construction of direct tests for GL(CL) forecast superiority. Define

$$G(x) = (F_2(x) - F_1(x))sgn(x),$$

where $sgn(x) = 1$, if $x \geq 0$, and $sgn(x) = -1$, if $x < 0$. Here, $F_i(x)$ denotes the cumulative distribution function (CDF) of u_i, and

$$C(x) = \int_{-\infty}^{x} (F_1(t) - F_2(t))dt 1_{\{x<0\}} + \int_{x}^{\infty} (F_2(t) - F_1(t))dt 1_{\{x\geq 0\}}$$

Assumption 7.1: $g(\cdot) : \Re \to \Re^+$ is continuously differentiable, except for finitely many points, with derivative $\nabla g(\cdot)$, such that $\nabla g(z) \leq 0$, $\forall z \leq 0$ and $\nabla g(z) \geq 0$, $\forall z \geq 0$.

PROPOSITION 7.1 (From Propositions 2.2 and 2.3 in [41]): With Assumption 7.1, $E(g(u_{1,t})) \leq E(g(u_{2,t}))$, $\forall g(\cdot) \in GL(\cdot)$, if and only if $G(x) \leq 0$, $\forall x \in \mathscr{X}$, where $\mathscr{X}$ is the union of the supports of all forecast errors. Further, if $\int_{-\infty}^{x}(F_1(t)-F_2(t))dt1_{\{x<0\}}$ and $\int_{x}^{\infty}(F_2(t)-F_1(t))dt1_{\{x\geq 0\}}$ are well defined for each $x \in \mathscr{X}$, then $E(g(u_{1,t})) \leq E(g(u_{2,t}))$, $\forall g(\cdot) \in CL(\cdot)$ if and only if $C(x) \leq 0$, $\forall x \in \mathscr{X}$.

The above proposition establishes a clear mapping between GL (CL) forecast superiority and first (second) order stochastic dominance. Intuitively, if we construct a graph that contains a plot of $G(x)$ against x. When $u_1 \succeq_G u_2$, we expect all points lie below or on the zero line. Similarly, if we construct a graph that contains a plot of $C(x)$ against x. When $u_1 \succeq_C u_2$, we expect all points lie below or on the zero line as well.

The hypotheses tested in JCS are:

$$H_0 : \max_{i=1,\ldots,m} E(g(u_{0,t+1}) - g(u_{i,t+1})) \leq 0$$

versus

$$H_A : \max_{i=1,\ldots,m} E(g(u_{0,t+1}) - g(u_{i,t+1})) > 0$$

Given Proposition 7.1, the above hypotheses can be restated as

$$\begin{aligned} H_0^{TG} = H_0^{TG-} \cap H_0^{TG+} : & \Big(\max_{i=1,\ldots,m} (F_0(x) - F_i(x)) \leq 0, \ \forall x \leq 0 \Big) \\ & \cap \Big(\max_{i=1,\ldots,m} (F_i(x) - F_0(x)) \leq 0, \ \forall x > 0 \Big) \end{aligned}$$

versus

$$\begin{aligned} H_A^{TG} = H_A^{TG-} \cup H_A^{TG+} : & \Big(\max_{i=1,\ldots,m} (F_0(x) - F_i(x)) > 0, \text{ for some } x \leq 0 \Big) \\ & \cup \Big(\max_{i=1,\ldots,m} (F_i(x) - F_0(x)) > 0, \text{ for some } x > 0 \Big) \end{aligned}$$

for the case of GL forecast superiority. Similarly, for the case of CL forecast superiority, we have that:

$$\begin{aligned} H_0^{TC} = H_0^{TC-} \cap H_0^{TC+} : & \Big(\max_{i=1,\ldots,m} \int_{-\infty}^{x} (F_0(x) - F_i(x)) \leq 0, \ \forall x \leq 0 \Big) \\ & \cap \Big(\max_{i=1,\ldots,m} \int_{x}^{\infty} (F_i(x) - F_0(x)) \leq 0, \ \forall x > 0 \Big) \end{aligned}$$

versus

$$H_A^{TC} = H_A^{TC-} \cup H_A^{TC+} : \big(\max_{i=1,\ldots,m} \int_{-\infty}^{x} (F_0(x) - F_i(x)) > 0, \text{ for some } x \leq 0 \big)$$
$$\cup \big(\max_{i=1,\ldots,m} \int_{x}^{\infty} (F_i(x) - F_0(x)) > 0, \text{ for some } x > 0 \big)$$

Of note is that the above null (alternative) is the intersection (union) of two different null (alternative) hypotheses because of a discontinuity at zero. The test statistics for GL forecast superiority are constructed as follows:

$$TG_n^+ = \max_{i=1,\ldots,k} \sup_{x \in \mathscr{X}^+} \sqrt{n} \widehat{G}_{i,n}(x)$$

and

$$TG_n^- = \max_{i=1,\ldots,k} \sup_{x \in \mathscr{X}^-} \sqrt{n} \widehat{G}_{i,n}(x)$$

with

$$\widehat{G}_{i,n}(x) = \big(\widehat{F}_{0,n}(x) - \widehat{F}_{i,n}(x) \big) sgn(x)$$

where $\widehat{F}_{i,n}(x)$ denotes the empirical CDF of u_i, with

$$\widehat{F}_{i,n}(x) = n^{-1} \sum_{t=R}^{T} 1_{\{u_{i,t} \leq x\}}$$

Similarly, the test statistics for CL forecast superiority are constructed as follows:

$$TC_n^+ = \max_{i=1,\ldots,k} \sup_{x \in \mathscr{X}^+} \sqrt{n} \widehat{C}_{i,n}(x)$$

and

$$TC_n^- = \max_{i=1,\ldots,k} \sup_{x \in \mathscr{X}^-} \sqrt{n} \widehat{C}_{i,n}(x)$$

with

$$\widehat{C}_{i,n}(x) = \int_{-\infty}^{x} \big(\widehat{F}_{0,n}(x) - \widehat{F}_{i,n}(x) \big) dx 1_{\{x<0\}} - \int_{x}^{\infty} \big(\widehat{F}_{i,n}(x) - \widehat{F}_{0,n}(x) \big) dx 1_{\{x \geq 0\}}$$
$$= \frac{1}{n} \sum_{t=1}^{n} \Big\{ [(u_{0,t} - x) sgn(x)]_+ - [(u_{i,t} - x) sgn(x)]_+ \Big\},$$

where $[z]_+ = \max\{0, z\}$.

Note that in order to reduce computation time, it may be preferable to construct approximations to the suprema in statistics TG^+, TG^-, TC^+ and TC^- by taking maxima over some smaller grid of points, $\mathscr{X}_N = \{x_1, \ldots, x_N\}$, where $N < n$. Theoretically, the distribution theory is unaffected by using this approximation, as the set of evaluation points becomes dense in the joint support. We now require the following assumptions.

Assumption 7.2: (i) $\{(y_t, Z_i^t)'\}$ is a strictly stationary and α-mixing sequence with mixing coefficient $\alpha(l) = O(l^{-C_0})$, for some $C_0 > \max\{(q-1)(q+1), 1 + 2/\delta\}$,

with $i = 0, ..., m$, where q is an even integer that satisfies $q > 3(g_{\max} + 1)/2$. Here, $g_{\max} = \max\{g_0, ..., g_m\}$ and δ is a positive constant;
(ii) For $i = 0, ..., m$, $f_i(Z_i^t, \theta_i)$ is differentiable a.s. with respect to θ_i in the neighborhood $\Theta_i^\dagger$ of $\theta_i^\dagger$, with $\sup_{\theta \in \Theta_0^\dagger} ||\nabla_\theta f_i(Z_i^t, \theta_i)||_2 < \infty$;
(iii) The conditional distribution of $u_{i,t}$ given Z_i^t has bounded density with respect to the Lebesgue measure a.s., and $||u_{i,t}||_{2+\delta} < \infty$, $\forall i$.

Assumption 7.2*: (i) $\{(y_t, Z_i^t)'\}$ is a strictly stationary and α-mixing sequence with mixing coefficient $\alpha(l) = O(l^{-C_0})$, for some $C_0 > \max\{rq/(r-q), 1 + 2/\delta\}$, with $i = 0, ..., m$, and $r > q > g_{\max} + 1$;
(ii) For $i = 0, ..., m$, $f_i(Z_i^t, \theta_i)$ is differentiable a.s. with respect to θ_i in the neighborhood $\Theta_i^\dagger$ of $\theta_i^\dagger$, with $\sup_{\theta \in \Theta_0^\dagger} ||\nabla_\theta f_i(Z_i^t, \theta_i)||_r < \infty$;
(iii) $||u_{i,t}||_r < \infty$, $\forall i$.

Assumption 7.3: $\forall\, i$ and t, $\widehat{\theta}_{i,t}$ satisfies $\widehat{\theta}_{i,t} - \theta_i^\dagger = B_i(t)H_i(t)$, where $B_i(t)$ is a $n_i \times L_i$ matrix and $H_i(t)$ is $L_i \times 1$, with the following:
(i) $B_i(t) \to B_i$ a.s., where B_i is a matrix of rank n_i;
(ii) $H_i(t) = t^{-1}\sum_{s=1}^t h_{i,s}$, $R^{-1}\sum_{s=t-R+1}^t h_{i,s}$ and $R^{-1}\sum_{s=1}^R h_{i,s}$ for the recursive, rolling and fixed schemes, respectively, where $h_{i,s} = h_{i,s}(\theta_i^\dagger)$;
(iii) $E(h_{i,s}(\theta_i^\dagger) = 0$; and
(iv) $||h_{i,s}(\theta_i^\dagger)||_{2+\delta} < \infty$, for some $\delta > 0$.

Assumption 7.4: (i) The distribution function of forecast errors, $F_i(x, \theta_i)$ is differentiable with respect to θ_i in a neighborhood $\Theta_i^\dagger$ of $\theta_i^\dagger$, $\forall i$;
(ii) $\forall i$, and $\forall$ sequences of positive constants $\{\xi_n : n \geq 1\}$, such that $\xi_n \to 0$, $\sup_{x \in \mathscr{X}} \sup_{\theta:\, ||\theta - \theta_i^\dagger|| \leq \xi_n} ||\nabla_\theta F_i(x, \theta) sgn(x) - \Delta_i^\dagger(x)|| = O(\xi_n^\eta)$, for some $\eta > 0$, where $\Delta_i^\dagger(x) = \nabla_\theta F_i(x, \theta_i^\dagger) sgn(x)$;
(iii) $\sup_{x \in \mathscr{X}} ||\Delta_i^\dagger(x)|| < \infty, \forall i$.

Assumption 7.4*: (i) Assumption 5.4 (i) holds;
(ii) $\forall i$, and $\forall$ sequences of positive constants $\{\xi_n : n \geq 1\}$, such that $\xi_n \to 0$, $\sup_{x \in \mathscr{X}} \sup_{\theta:\, ||\theta - \theta_i^\dagger|| \leq \xi_n} ||\nabla_\theta \{\int_{-\infty}^x F_i(t, \theta) dt 1_{\{x<0\}} + \int_x^\infty (1 - F_i(x, \theta)) dt 1_{\{x \geq 0\}}\} - \Lambda_i^\dagger(x)|| = O(\xi_n^\eta)$, for some $\eta > 0$, where

$$\Lambda_i^\dagger(x) = \nabla_\theta \{\int_{-\infty}^x F_i(t, \theta_i^\dagger) dt 1_{\{x<0\}} + \int_x^\infty (1 - F_i(x, \theta_i^\dagger)) dt 1_{\{x \geq 0\}}\};$$

(iii) $\sup_{x \in \mathscr{X}} ||\Lambda_i^\dagger(x)|| < \infty, \forall i$.

Assumptions 7.2* and 7.4* are needed for testing H_0^{TC}. Note that the first and third assumptions parallel those imposed by [42], with the uniform continuity conditions in Assumptions 7.4 and 7.4* strengthened. Assumption 7.2 is needed in order to verify the stochastic equicontinuity of the empirical process, for a class of

bounded functions that appears in the TG_n test. Assumption 7.2* introduces a trade-off between mixing sizes and moment conditions, and is used to verify the stochastic equicontinuity result for the possibly unbounded functions that appear in the TC_n test. For further details, see [45]. Assumptions 7.4 and 7.4* differ in the amount of smoothness required. For the CL forecast superiority test, less smoothness is required. Finally, it is worth stressing that Assumptions 7.3 and 7.5 are identical to Assumptions 1 and 2 in [46], respectively.

PROPOSITION 7.2 (From Theorem 3.1 in [41]): (i) With Assumptions 4.3, 7.2–7.4, under $H_0^{TG^-}$,

$$TG_n^- \xrightarrow{d} \max_{i=1,\ldots,m} \sup_{x\in\mathscr{B}_i^{g-}} [\widetilde{g}_i(x) + \Delta_{i0}(x)' B_i \upsilon_{i0} - \Delta_{10}(x)' B_1 \upsilon_{10}], \text{ if } TG^- = 0$$
$$\to -\infty, \text{ if } TG^- < 0$$

Under H_0^{TG+},

$$TG_n^+ \xrightarrow{d} \max_{i=1,\ldots,m} \sup_{x\in\mathscr{B}_i^{g+}} [\widetilde{g}_i(x) + \Delta_{i0}(x)' B_i \upsilon_{i0} - \Delta_{10}(x)' B_1 \upsilon_{10}], \text{ if } TG^+ = 0$$
$$\to -\infty, \text{ if } TG^+ < 0$$

where $\mathscr{B}_i^{g-} = \{x \in \mathscr{X}^- : F_0(x) = F_i(x)\}$ and $\mathscr{B}_i^{g+} = \{x \in \mathscr{X}^+ : F_0(x) = F_i(x)\}$, and $\left(\widetilde{g}_i(\cdot), \upsilon_{i0}, \upsilon_{10}\right)'$ is a mean zero Gaussian process with covariance function given by

$$\Omega_i^g(x_1,x_2) = \lim_{T\to\infty} E \begin{pmatrix} \upsilon_{i,n}^g(x_1,\theta_i^\dagger) - \upsilon_{0,n}^g(x_1,\theta_0^\dagger) \\ \sqrt{n}\overline{H}_{i,n} \\ \sqrt{n}\overline{H}_{0,n} \end{pmatrix} \begin{pmatrix} \upsilon_{i,n}^g(x_2,\theta_i^\dagger) - \upsilon_{0,n}^g(x_2,\theta_0^\dagger) \\ \sqrt{n}\overline{H}_{i,n} \\ \sqrt{n}\overline{H}_{0,n} \end{pmatrix}'$$

with $\overline{H}_{i,n} = n^{-1}\sum_{t=R}^T H_i(t)$, and $\upsilon_{i,n}^g(x,\theta)$ is an empirical process defined as

$$\upsilon_{i,n}^g(x,\theta) = \frac{1}{\sqrt{n}} \sum_{t=R}^T \{1_{\{u_{i,t+\tau}(\theta)\le x\}} - F_i(x,\theta)\} sgn(x)$$

(ii) With Assumptions 7.2*, 7.3, 7.4* and 7.5, under $H_0^{TC^-}$,

$$TC_n^- \xrightarrow{d} \max_{i=1,\ldots,m} \sup_{x\in\mathscr{B}_i^{c-}} [\widetilde{c}_i(x) + \Lambda_{i0}(x)' B_i \upsilon_{i0} - \Lambda_{10}(x)' B_1 \upsilon_{10}], \text{ if } TC^- = 0$$
$$\to -\infty, \text{ if } TC^- < 0$$

Under $H_0^{TC^+}$,

$$TC_n^+ \xrightarrow{d} \max_{i=1,\ldots,m} \sup_{x\in\mathscr{B}_i^{c+}} [\widetilde{c}_i(x) + \Lambda_{i0}(x)' B_i \upsilon_{i0} - \Lambda_{10}(x)' B_1 \upsilon_{10}], \text{ if } TC^+ = 0$$
$$\to -\infty, \text{ if } TC^+ < 0$$

where $\mathscr{B}_i^{c-} = \{x \in \mathscr{X}^- : \int_{-\infty}^{x} (F_i(x) - F_0(x))dx 1_{\{x<0\}}\}$ and $\mathscr{B}_i^{c+} = \{x \in \mathscr{X}^+ : \int_{x}^{\infty} (F_0(x) - F_i(x))dx 1_{\{x\geq 0\}}\}$. Similarly, $\left(\widetilde{c}_i(\cdot), \upsilon_{i0}, \upsilon_{10}\right)'$ is a mean zero Gaussian process with covariance function given by

$$\Omega_i^c(x_1,x_2) = \lim_{T\to\infty} E \begin{pmatrix} \upsilon_{i,n}^c(x_1,\theta_i^\dagger) - \upsilon_{0,n}^c(x_1,\theta_0^\dagger) \\ \sqrt{nH}_{i,n} \\ \sqrt{nH}_{0,n} \end{pmatrix} \begin{pmatrix} \upsilon_{i,n}^c(x_2,\theta_i^\dagger) - \upsilon_{0,n}^c(x_2,\theta_0^\dagger) \\ \sqrt{nH}_{i,n} \\ \sqrt{nH}_{0,n} \end{pmatrix}'$$

where $\upsilon_{i,n}^c(x,\theta)$ is an empirical process defined as

$$\upsilon_{i,n}^c(x,\theta) = \frac{1}{\sqrt{n}} \sum_{t=R}^{T} \Big\{ \int_{-\infty}^{x} [1_{\{u_{i,t+\tau}(\theta)\leq s\}} - F_i(s,\theta)]ds 1_{\{x<0\}} - \int_{x}^{\infty} [1_{\{u_{i,t+\tau}(\theta)\leq s\}} - F_i(s,\theta)]ds 1_{\{x\geq 0\}} \Big\}$$

The asymptotic null distributions of TG_n^+ (TG_n^-) and TC_n^+ (TC_n^-) depend on the true model parameters and the distribution functions, $F_i(\cdot),\ i = 1,...,m$, which implies that the asymptotic critical values for TG_n^+ (TG_n^-) and TC_n^+ (TC_n^-) cannot be tabulated. Therefore, the stationary bootstrap is used to approximate the asymptotic null distributions of our test statistics. (Note that the block bootstrap can also be used, as discussed in subsequent research by Corradi, Sin and Swanson.) The objective is to utilize bootstrap procedure that mimics the asymptotic null distribution in the least favorable case, where $F_0(x) = ... = F_m(x)$, $\forall x \in \mathscr{X}$.

Define the bootstrap statistic as:

$$TG_n^{*+} = \max_{i=1,\ldots,k} \sup_{x\in\mathscr{X}^+} \sqrt{n}\Big(\widehat{G}_{i,n}^*(x) - \widehat{G}_{i,n}(x)\Big)$$

with

$$\widehat{G}_{i,n}^*(x) = \big(\widehat{F}_{0,n}^*(x) - \widehat{F}_{i,n}^*(x)\big) sgn(x)$$

where $\widehat{F}_{i,n}^*(x)$ denotes the empirical CDF of resampled u_i, i.e. u_i^*. TG_n^{*-}, TC_n^{*+} and TC_n^{*-} can be defined analogously.

Assumption 7.6: The smoothing parameter, S_n, determining the mean block length in stationary bootstrap satisfies $0 < S_n < 1$, $S_n \to 0$ and $nS_n^2 \to \infty$, as $n \to \infty$.

Assumption 7.7: For any arbitrary $n_i \times 1$ vector, λ_i, with $\lambda_i^{'}\lambda_i = 1$, and $\forall i$, we have (i)

$$\Pr\Big[\limsup_{t\geq R} n^{1/2}\frac{|\lambda_i'(\widehat{\theta}_{i,t}-\theta_i^{\dagger})|}{(\lambda_i'\Sigma_i\lambda_i \text{loglog}(\lambda_i'\Sigma_i\lambda_i)n)^{1/2}}=1\Big]=1$$

for the recursive scheme, where $\Sigma_i = B_i[\lim_{T\to\infty}\text{Var}(n^{-1/2}\sum_{t=R+1}^{T}H_i(t))]B_i'$.
(ii)

$$\Pr\Big[\limsup_{t\geq R} R^{1/2}\frac{|\lambda_i'(\widehat{\theta}_{i,t}-\theta_i^{\dagger})|}{(\lambda_i'\Sigma_i\lambda_i \text{loglog}(\lambda_i'\Sigma_i\lambda_i)R)^{1/2}}=1\Big]=1$$

for the rolling scheme, where $\Sigma_i = B_i[\lim_{T\to\infty}\text{Var}(R^{-1/2}\sum_{t=R+1}^{T}H_i(t))]B_i'$.

PROPOSITION 7.3 (From Corollary 3.3 in [41]): With Assumptions 7.2–7.4, 7.6 and 7.7, and that $(n/R)\text{loglog}R\to 0$, as $T\to\infty$, then

$$\rho\Big(L[\max_{i=1,...,m}\sup_{x\in\mathscr{X}^+}\sqrt{n}(\widehat{G}^*_{i,n}(x)-\widehat{G}_{i,n}(x))|U_1,...,U_{T+\tau}],$$
$$L[\max_{i=1,...,m}\sup_{x\in\mathscr{X}^+}\sqrt{n}(\widehat{G}_{i,n}(x)-G_i(x))]\Big)\xrightarrow{n}0$$

and

$$\rho\Big(L[\max_{i=1,...,m}\sup_{x\in\mathscr{X}^-}\sqrt{n}(\widehat{G}^*_{i,n}(x)-\widehat{G}_{i,n}(x))|U_1,...,U_{T+\tau}],$$
$$L[\max_{i=1,...,m}\sup_{x\in\mathscr{X}^-}\sqrt{n}(\widehat{G}_{i,n}(x)-G_i(x))]\Big)\xrightarrow{n}0$$

where ρ is any metric metrizing weak convergence, $L[\cdot]$ denotes the probability law of the corresponding Hilbert space valued random variable, and $U_t = (y_t, Z_0^t, ..., Z_m^t)'$.

Therefore, the asymptotic null distribution of TG_n^+ (TG_n^-) can be approximated by $TG_n^{*+} - TG_n^+$ ($TG_n^{*-} - TG_n^-$). Arguments in favor of using the stationary bootstrap with TC_n^+ and TC_n^- are similar.

To conduct inference, use the following approach due to [44]. Define $q^{G^+}_{n,S_n}(1-\alpha)$ and $q^{G^-}_{n,S_n}(1-\alpha)$ to be the $(1-\alpha)$-th sample quantile of TG_n^{*+} and TG_n^{*-}, respectively. Then, estimate bootstrap p-values, $p^{G^+}_{B,n,S_n} = \frac{1}{B}\sum_{s=1}^{B}(TG_n^{*+}\geq TG_n^+)$, and finally use the following rules.

Rule TG: Reject H_0^{TG} at level α, if $\min\Big\{p^{G^+}_{B,n,S_n},\ p^{G^-}_{B,n,S_n}\Big\}\leq\alpha/2$;
Rule TC: Reject H_0^{TG} at level α, if $\min\Big\{p^{C^+}_{B,n,S_n},\ p^{C^-}_{B,n,S_n}\Big\}\leq\alpha/2$;

Note that Holm bounds are equivalent to Bonferroni bounds when there are only two hypotheses. From Proposition 7.3, it follows immediately that this test,when implemented using the stationary bootstrap, has asymptotically correct size only in the least favorable case, under the null, and is asymptotically biased towards certain

local alternatives.

PROPOSITION 7.4 (From Theorem 4.1 in [41]): With Assumptions 4.3, 7.2–7.4, under H_A^{TG},

$$\Pr(TG_n^+ > q_{n,S_n}^{G^+}(1-\alpha)) \to 1, \ as \ T \to \infty$$

and

$$\Pr(TG_n^- > q_{n,S_n}^{G^-}(1-\alpha)) \to 1, \ as \ T \to \infty$$

The above proposition ensures unit asymptotic power under the alternative. Similar arguments apply to TC_n^+ and TC_n^- as well. For details of the power of TG_n^+ (TG_n^-) and TC_n^+ (TC_n^-) tests against a sequence of contiguous local alternatives converging to the null, at rate $n^{-1/2}$, see [41].

References

1. Diebold, F.X. and Mariano, R.S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20(1), 134–144.
2. West, K.D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64, 1067–1084
3. Corradi, V., Swanson, N.R. and Olivetti, C. (2001). Predictive ability with cointegrated variables. *Journal of Econometrics*, 104(2), 315–358.
4. Rossi, B. (2005). Testing long-horizon predictive ability with high persistence, and the MeeseRogoff puzzle. *International Economic Review*, 46(1), 61–92.
5. Meese, R.A. and Rogoff, K. (1983). Empirical exchange rate models of the seventies: Do they fit out-of-sample? *Journal of International Economics*, 14, 3–24.
6. Clark, T.E. and McCracken, M.W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105(1), 85–110.
7. Clark, T.E. and McCracken, M.W. (2003). Evaluating long horizon forecasts. *Working Paper, University of Missouri-Columbia*.
8. Kilian, L. (1999). Exchange rates and monetary fundamentals: What do we learn from longhorizon regressions? *Journal of Applied Econometrics*, 14(5), 491–510.
9. Chao, J., Corradi, V. and Swanson, N.R. (2001). Out-of-sample tests for Granger causality. *Macroeconomic Dynamics*, 5(4), 598–620.
10. White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5), 1097–1126.
11. Corradi, V. and Swanson, N. R. (2007). Nonparametric bootstrap procedures for predictive inference based on recursive estimation schemes. *International Economic Review*, 48(1), 67–109.
12. Corradi, V. and Swanson, N. R. (2006). Predictive density and conditional confidence interval accuracy tests. *Journal of Econometrics*, 135(1), 187–228.
13. Hansen, R.P. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4), 365–380.
14. Corradi, V. and Distaso, W. (2011). Multiple forecast model evaluation. *The Oxford Handbook of Economic Forecasting, Oxford University Press, USA*, 391–414.
15. Andrews, D.W. and Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78(1), 119–157.
16. Romano, J.P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), 1237–1282.
17. Politis, D.N., Romano, J.P. and Wolf, M. (1999). Subsampling. *Springer Series in Statistics. New York*.
18. Bierens, H.J. (1990). A consistent conditional moment test of functional form. *Econometrica*, 58, 1443–1458.
19. Bierens, H.J. and Ploberger, W. (1997). Asymptotic theory of integrated conditional moment tests. *Econometrica*, 65, 1129–1151.
20. De Jong, R.M. (1996). The Bierens test under data dependence. *Journal of Econometrics*, 72(1), 1–32.
21. Hansen, B.E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, 64, 413–430.
22. Lee, T.H., White, H. and Granger, C.W.J. (1993). Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests. *Journal of Econometrics*, 56(3), 269–290.
23. Stinchcombe, M.B. and White, H. (1998). Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory*, 14(3), 295–325.
24. Corradi, V. and Swanson, N.R. (2002). A consistent test for out of sample nonlinear predictive ability. *Journal of Econometrics*, 110, 353–381.
25. Granger, C.W.J. (1993). On the limitations of comparing mean square forecast errors: A comment. *Journal of Forecasting*, 12(8), 651–652.

26. Weiss, A. (1996). Estimating time series models using the relevant cost function. *Journal of Applied Econometrics*, 11(5), 539–560.
27. White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1), 1–25.
28. Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2), 307–333.
29. Gianni, A. and Giacomini R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, 25(2), 177–190.
30. Kitamura, Y. (2002). Econometric comparisons of conditional models. *Working Paper, University of Pennsylvania*.
31. Schorfheide, F. (2010). Loss function-based evaluation of DSGE models. *Journal of Applied Econometrics*, 15(6), 645–670.
32. Fernández-Villaverde, J. and Rubio-RamíRez, J. F. (2004). Comparing dynamic equilibrium models to data: A bayesian approach. *Journal of Econometrics*, 123(1), 153–187.
33. Chang, Y., Gomes, J. F. and Schorfheide, F. (2002). Learning-by-doing as a propagation mechanism. *American Economic Review*, 92(5), 1498–1520.
34. Corradi, V. and Swanson, N.R. (2005). A test for comparing multiple misspecified conditional interval models. *Econometric Theory*, 21(5), 991–1016.
35. Corradi, V. and Swanson, N.R. (2006). Predictive density and conditional confidence interval accuracy tests. *Journal of Econometrics*, 135(1), 187–228.
36. Hall, P. and Horowitz, J.L. (1996). Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica*, 64(4), 891–916.
37. Andrews, D.W.K. (2002). Higher-order improvements of a computationally attractive k-step bootstrap for extremum estimators. *Econometrica*, 70(1), 119–162.
38. Andrews, D.W.K. (2004). The blockblock bootstrap: Improved asymptotic refinements. *Econometrica*, 72(3), 673-700.
39. Inoue, A. and Shintani, M. (2006). Bootstrapping GMM estimators for time series. *Journal of Econometrics*, 133(2), 531–555.
40. Linton, O.B., Maasoumi, E. and Whang, Y.J. (2002). Consistent testing for stochastic dominance: A subsampling approach. *Social Science Electronic Publishing*, 72(3), 735–765.
41. Jin, S., Corradi V. and Swanson, N.R. (2017). Robust forecast comparison. *Econometric Theory*, 33(6), 1306–1351.
42. Linton, O., Maassoumi E. and Whang Y.J. (2005). Consistent tesing for stochastic dominance: A subsampling approach. *Review of Economic Studies*, 72, 735–765.
43. Granger, C.W.J (1999). Outline of forecast theory using generalized cost function. *Spanish Economic Review*, 1, 161–173.
44. Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70.
45. Hansen, B.E. (1996) Stochastic equicontinuity for unbounded dependent heterogeneous arrays. *Econometric Theory*, 12, 347–359.
46. McCracken, M.W. (2000) Robust out-of-sample inference. *Journal of Econometrics*, 99, 195–223.